ESTADÍSTICA APLICADA

N. GUARÍN S.
nguarins@epm.net.co

Tabla de Contenido

Introducción

1. La Estadística

1.1 Importancia
1.2 Definición
1.3 División

2. Etapas del Método Estadístico

2.1 Planteamiento del problema
2.2 Fijación de los objetivos
2.3 Formulación de las hipótesis
2.4 Definición de la unidad de observación y de la unidad de medida
2.5 Determinación de la población y de la muestra
2.6 La recolección
2.7 Crítica, clasificación y ordenación
2.8 La tabulación
2.9 La presentación
2.10 El análisis
2.11 Publicación
Cuestionario

3. Distribución de Frecuencias

3.1 Distribución de frecuencias simple
Ejercicios
3.2 Distribución de frecuencias por intervalo
3.3 Reglas empíricas para la construcción de Intervalos
Cuestionario y ejercicios propuestos

4. Representación Gráfica

4.1 Definición
4.2 Componentes de una gráfica

4.3 Principales tipos de gráficos
4.3.1 Gráfico de líneas
4.3.2 Gráfico de líneas compuesto
4.3.3 Gráfico de barras
4.3.4 Gráfico de barras compuesto
4.3.5 Gráfico de sectores circulares
4.3.6 Histograma de frecuencias
4.3.7 Polígono de frecuencias
4.3.8 Histograma de frecuencias acumuladas
Cuestionario y ejercicios propuestos

5. **Medidas de Tendencia Central**

5.1 Media aritmética
5.1.1 Propiedades de la media aritmética
5.1.2 Media aritmética con cambio origen y de escala
5.1.3 Media aritmética ponderada
5.2 Mediana
5.2.1 La mediana cuando los datos no están agrupados en intervalos
5.2.2 La mediana cuando la información está agrupada en intervalos
5.3 La Moda
5.3.1 La moda cuando los datos no están agrupados en intervalos
5.3.2 Cálculo de la moda con la información agrupada en intervalos
Cuestionario y ejercicios propuestos

6. **Medidas de Posición (Percentiles)**

6.1 Cuartiles
6.2 Quintiles
6.3 Deciles
6.4 Centiles
6.5 Resumen
Cuestionario y ejercicios propuestos

7. **Medidas de Dispersión**

7.1 Rango o recorrido
7.2 Desviación media
7.3 Varianza
7.4 Coeficiente de variabilidad
Cuestionario y ejercicios propuestos

8. Regresión y Correlación Lineal

8.1 Tablas de doble entrada
8.2 Correlación
8.3 Regresión lineal
8.3.1 Ajuste rectilíneo (método de los mínimos cuadrados)
8.3.2 Ajuste parabólica (método de los mínimos cuadrados)
Cuestionario y ejercicios propuestos

9. Tasas e Índices

9.1 Tasa
9.2 Índice
9.2.1 Índice simple
9.2.1.1 Índice de base fija
9.2.1.2 Índice de base móvil
9.2.2 Índices compuestos (globales)
9.2.2.1 Índice de Laspeyres
9.2.2.2 Índice de Passche
9.2.2.3 Índice ideal de Fisher
Cuestionario y ejercios propuestos

10. Nociones de Probabilidad (Eventos)

10.1 Nociones de conteo
10.1.1 Principio fundamental 1
10.1.2 Principio fundamental 2
10.1.3 Permutaciones
10.1.4 Variaciones
10.1.5 Combinaciones
10.1.6 Permutaciones con repetición
10.1.7 Variaciones con repetición
Ejercicios propuestos
10.2 Definición de probabilidad
10.2.1 Probabilidad a priori
10.2.2 Probabilidad a posteriori
10.2.3 Probabilidad subjetiva
10.3 Axiomas de la teoría de probabilidades
10.4 Probabilidad condicional e independencia estadística
Cuestionario y ejercicios propuestos

10.5 Variable aleatoria
10.6 Función de probabilidad
10.6.1 Función de probabilidad
10.6.2 Función de distribución
Cuestionario y ejercicios propuestos
10.7 Valor esperado (esperanza matemática)
10.7.1 Media aritmética poblacional
10.7.2 Varianza poblacional
Cuestionario y ejercicios propuestos

11. Distribuciones Especiales

11.1 Distribución de Bernoulli
11.2 Distribución binomial, tablas binomiales
11.3 Distribución de Poisson, tablas de Poisson
11.4 Distribución normal
11.5 Distribución normal estándar, tablas normales
Cuestionario y ejercicios propuestos
11.6 El tamaño de la muestra

Apéndice No. 1

Apéndice No. 2

Apéndice No. 3

Solución a Algunos Ejercicios Propuestos

Enlaces

Bibliografía

Introducción

Las acciones que acometemos hoy
se basan en un plan de ayer y
las expectativas del mañana.

La palabra estadística se origina, en las técnicas de recolección, organización, conservación, y tratamiento de los datos propios de un estado, con que los antiguos gobernantes controlaban sus súbditos y dominios económicos. Estas técnicas evolucionaron a la par con el desarrollo de las matemáticas, utilizando sus herramientas en el proceso del análisis e interpretación de la información.

Para mediados del siglo XVII en Europa, los juegos de azar eran frecuentes, aunque sin mayores restricciones legales. El febril jugador De Méré consultó al famoso matemático y filosofo Blaise Pascal (1623-1662) para que le revelara las leyes que controlan el juego de los dados, el cual, interesado en el tema, sostuvo una correspondencia epistolar con el tímido Pierre de Fermat (1601-1665, funcionario público apasionado por las matemáticas; célebre porque no publicaba sus hallazgos) dando origen a la teoría de la probabilidad, la cual se ha venido desarrollando y constituyéndose en la base primordial de la estadística.

En nuestros días, son de uso cotidiano las diferentes técnicas estadísticas que partiendo de observaciones muestrales o históricas, crean modelos lógico-matemáticos que se "aventuran" describir o pronosticar un determinado fenómeno con cierto grado de certidumbre medible.

El presente texto no pretende teorizar el saber estadístico, desde luego, no es un libro para estadísticos, ya que, adrede se obvia el rigor científico de lo expuesto en beneficio de la sencillez necesaria para el neófito; con un lenguaje coloquial se conduce al lector a través del contenido, a partir de dos o tres ejemplos que ilustran la aplicabilidad de los temas tratados.

El avance tecnológico en la informática ha contribuido enormemente al desarrollo de la estadística, sobre todo en la manipulación de la información, pues en el mercado existen paquetes estadísticos de excelente calidad, como el SAS, SPSS, SCA, STATGRAPHICS, amén de otros, que "corren" en un ordenador sin mayores exigencias técnicas, permitiendo el manejo de grandes volúmenes de información y de variables.

La estadística, entonces, dejó de ser una técnica exclusiva de los estados, para convertirse en una herramienta imprescindible de todas las ciencias, de donde proviene la desconcertante des-uniformidad en las definiciones de los diferentes autores, ya que cada estudioso la define de acuerdo con lo que utiliza de ella y tenemos definiciones como que: la estadística es la tecnología del método científico, o que es el conocimiento relacionado

con la toma de decisiones en condiciones de incertidumbre, o que la estadística son métodos para obtener conclusiones a partir de los resultados de los experimentos o procesos, o que es un método para describir o medir las propiedades de una población. En fin, no se trata de discutir si la estadística es una ciencia, una técnica o una herramienta, sino de la utilización de sus métodos en provecho de la evolución del conocimiento.

La estadística hace inferencias sobre una población, partiendo de una muestra representativa de ella. Es a partir del proceso del diseño y toma de la muestra desde donde comienzan a definirse las bondades y confiabilidad de nuestras aseveraciones, hechas, preferentemente, con un mínimo costo y mínimo error posible.

1. La Estadística

" El poder se nutre de la información
y el conocimiento".

1.1 IMPORTANCIA

En las últimas décadas la estadística ha alcanzado un alto grado de desarrollo, hasta el punto de incursionar en la totalidad de las ciencias; inclusive, en la lingüística se aplican técnicas estadísticas para esclarecer la paternidad de un escrito o los caracteres más relevantes de un idioma.

La estadística es una ciencia auxiliar para todas las ramas del saber; su utilidad se entiende mejor si tenemos en cuenta que los quehaceres y decisiones diarias embargan cierto grado de incertidumbre... y la Estadística ayuda en la incertidumbre, trabaja con ella y nos orienta para tomar las decisiones con un determinado grado de confianza.

Los críticos de la estadística afirman que a través de ella es posible probar cualquier cosa, lo cual es un concepto profano que se deriva de la ignorancia en este campo y de lo polifacético de los métodos estadísticos. Sin embargo muchos "investigadores" tendenciosos han cometido abusos con la estadística, elaborando "investigaciones" de intención, teniendo previamente los resultados que les interesan mostrar a personas ingenuas y desconocedoras de los hechos. Otros, por ignorancia o negligencia, abusan de la estadística utilizando modelos inapropiados o razonamientos ilógicos y erróneos que conducen al rotundo fracaso de sus investigaciones.

Lincoln L. Chao* hace referencia a uno de los más estruendosos fracasos, debido a los abusos en la toma de una muestra:

Se trata del error cometido por la Literary Digest que, en sus pronósticos para las elecciones presidenciales en EE.UU. para 1936, afirmó que Franklin D. Roosvelt obtendría 161 votos electorales y Alfred Landon, 370. La realidad mostró a Roosvelt con 523 votos y a Landon con 8 solamente.

El error se debió a que la muestra fue tomada telefónicamente a partir de la lista de suscriptores de la Digest y, en 1936, las personas que se daban el lujo de tener teléfonos y suscripciones a revistas no configuraban una muestra representativa de los votantes de EE.UU. y, por ende, no podía hacerse un pronóstico confiable con tan sesgada información.

1.2 DEFINICIÓN

Definir la estadística es una tarea difícil porque tendríamos que definir cada una de las técnicas que se emplean en los diferentes campos en los que interviene. Sin embargo, diremos, en forma general, que la estadística es un conjunto de técnicas que, partiendo de la observación de fenómenos, permiten al investigador obtener conclusiones útiles sobre ellos.

1.3 DIVISIÓN

La estadística se divide en dos grandes ramas de estudio que son: ***La estadística descriptiva***, la cual se encarga de la recolección, clasificación y descripción de datos muestrales o poblacionales, para su interpretación y análisis, que es de la que nos ocuparemos en este curso; y ***la estadística matemática o inferencial***, que desarrolla modelos teóricos que se ajusten a una determinada realidad con cierto grado de confianza.

Estas dos ramas no son independientes; por el contrario, son complementarias y entre ambas dan la suficiente ilustración sobre una posible realidad futura, con el fin de que quien tenga poder de decisión, tome las medidas necesarias para transformar ese futuro o para mantener las condiciones existentes.

---

* LINCOLN, L. Chao. Estadística para Ciencias Administrativas. Trad. Jesús María Castaño.

2. Etapas del Método Estadístico

El método estadístico, parte de la observación de un fenómeno, y como no puede siempre mantener las mismas condiciones predeterminadas o a voluntad del investigador, deja que actúen libremente, pero se registran las diferentes observaciones y se analizan sus variaciones.

Para el planeamiento de una investigación, por norma general, se siguen las siguientes etapas:

2.1 Planteamiento del problema.
2.2 Fijación de los objetivos.
2.3 Formulación de la hipótesis.
2.4 Definición de la unidad de observación y de la unidad de medida.
2.5 Determinación de la población y de la muestra.
2.6 La recolección.
2.7 Crítica, clasificación y ordenación.
2.8 Tabulación.
2.9 Presentación.
2.10 Análisis.
2.11 Publicación.

2.1 PLANTEAMIENTO DEL PROBLEMA

Al abordar una investigación se debe tener bien definido qué se va a investigar y por qué se pretende estudiar algo. Es decir, se debe establecer una delimitación clara, concreta e inteligible sobre el o los fenómenos que se pretenden estudiar, para lo cual se deben tener en cuenta, entre otras cosas, la revisión bibliográfica del tema, para ver su accesibilidad y consultar los resultados obtenidos por investigaciones similares, someter nuestras proposiciones básicas a un análisis lógico; es decir, se debe hacer una ubicación histórica y teórica del problema.

2.2 FIJACIÓN DE LOS OBJETIVOS

Luego de tener claro lo que se pretende investigar, Debemos presupuestar hasta dónde queremos llegar; en otras palabras, debemos fijar cuales son nuestras metas y objetivos. Estos deben plantearse de tal forma que no haya lugar a confusiones o ambigüedades y

debe, además, establecerse diferenciación entre lo de corto, mediano y largo plazo, así como entre los objetivos generales y los específicos.

2.3 FORMULACIÓN DE LAS HIPÓTESIS

Una hipótesis es ante todo, una explicación provisional de los hechos objeto de estudio, y su formulación depende del conocimiento que el investigador posea sobre la población investigada. Una hipótesis estadística debe ser susceptible de docimar, esto es, debe poderse probar para su aceptación o rechazo.

Una hipótesis que se formula acerca de un parámetro (media, proporción, varianza, etc.), con el propósito de rechazarla, se llama Hipótesis de Nulidad y se representa por Ho; a su hipótesis contraria se le llama Hipótesis Alternativa (H1).

2.4 DEFINICIÓN DE LA UNIDAD DE OBSERVACIÓN Y DE LA UNIDAD DE MEDIDA

La Unidad de Observación, entendida como cada uno de los elementos constituyentes de la población estudiada, debe definirse previamente, resaltando todas sus características; pues, al fin de cuentas, es a ellas a las que se les hará la medición.

La unidad de observación puede estar constituida por uno o varios individuos u objetos y denominarse respectivamente simple o compleja.

El criterio sobre la unidad de medición debe ser previamente definido y unificado por todo el equipo de investigación. Si se trata de medidas de longitud, volumen, peso, etc., debe establecerse bajo qué unidad se tomarán las observaciones ya sea en metros, pulgadas, libras, kilogramos, etc.

Asociado a la unidad de medida, deben establecerse los criterios sobre las condiciones en las cuales se ha de efectuar la toma de la información.

2.5 DETERMINACIÓN DE LA POBLACIÓN Y DE LA MUESTRA

Estadísticamente, ***la población*** se define como un conjunto de individuos o de objetos que poseen una o varias características comunes. No se refiere esta definición únicamente a los seres vivientes; una población puede estar constituida por los habitantes de un país o por los peces de un estanque, así como por los establecimientos comerciales de un barrio o las unidades de vivienda de una ciudad.

Existen desde el punto de vista de su manejabilidad poblaciones finitas e infinitas. Aquí el término infinito no está siendo tomado con el rigor semántico de la palabra; por ejemplo, los peces dentro de un estanque son un conjunto finito; sin embargo, en términos estadísticos, puede ser considerado como infinito.

Muestra es un subconjunto de la población a la cual se le efectúa la medición con el fin

de estudiar las propiedades del conjunto del cual es obtenida.

En la práctica, estudiar todos y cada uno de los elementos que conforman la población no es aconsejable, ya sea por la poca disponibilidad de recursos, por la homogeneidad de sus elementos, porque a veces es necesario destruir lo que se está midiendo, por ser demasiado grande el número de sus componentes o no se pueden controlar; por eso se recurre al análisis de los elementos de una muestra con el fin de hacer inferencias respecto al total de la población. Existen diversos métodos para calcular el tamaño de la muestra y también para tomar los elementos que la conforman, pero no es el objetivo de este curso estudiarlos. Diremos solamente que la muestra debe ser representativa de la población y sus elementos escogidos al azar para asegurar la objetividad de la investigación.

. . . . .

2.6 LA RECOLECCIÓN

Una de las etapas más importantes de la investigación es la recolección de la información, la cual ha de partir, a menos que se tenga experiencia con muestras análogas, de una o varias muestras piloto en las cuales se pondrán a prueba los cuestionarios y se obtendrá una aproximación de la variabilidad de la población, con el fin de calcular el tamaño exacto de la muestra que conduzca a una estimación de los parámetros con la precisión establecida.

El establecimiento de las fuentes y cauces de información, así como la cantidad y complejidad de las preguntas, de acuerdo con los objetivos de la investigación son decisiones que se han de tomar teniendo en cuenta la disponibilidad de los recursos financieros, humanos y de tiempo y las limitaciones que se tengan en la zona geográfica, el grado de desarrollo, la ausencia de técnica, etc.

Es, entonces, descubrir dónde está la información y cómo y a qué "costo" se puede conseguir; es determinar si la encuesta se debe aplicar por teléfono, por correo, o si se necesitan agentes directos que recojan la información; establecer su número óptimo y preparar su entrenamiento adecuado.

2.7 CRITICA, CLASIFICACIÓN Y ORDENACIÓN

Después de haber reunido toda la información pertinente, se necesita la depuración de los datos recogidos. Para hacer la crítica de una información, es fundamental el conocimiento de la población por parte de quien depura para poder detectar falsedades en las respuestas, incomprensión a las preguntas, respuestas al margen, amén de todas las posibles causas de nulidad de una pregunta o nulidad de todo un cuestionario.

Separado el material de "desecho" con la información depurada se procede a establecer las clasificaciones respectivas y con la ayuda de hojas de trabajo, en las que se establecen los cruces necesarios entre las preguntas, se ordenan las respuestas y se preparan los modelos de tabulación de las diferentes variables que intervienen en la investigación.

El avance tecnológico y la popularización de los computadores hacen que estas tareas, manualmente dispendiosas, puedan ser realizadas en corto tiempo.

2.8 LA TABULACIÓN

Una ***tabla*** es un resumen de información respecto a una o más variables, que ofrece claridad al lector sobre lo que se pretende describir; para su fácil interpretación una tabla debe tener por lo menos: Un titulo adecuado el cual debe ser claro y conciso. La Tabla propiamente dicha con los correspondientes subtítulos internos y la cuantificación de los diferentes ítems de las variables, y las notas de pie de cuadro que hagan claridad sobre situaciones especiales de la tabla, u otorguen los créditos a la fuente de la información.

2.9 LA PRESENTACIÓN

Una información estadística adquiere más claridad cuando se presenta en la forma adecuada. Los cuadros, tablas y gráficos facilitan el análisis, pero se debe tener cuidado con las variables que se van a presentar y la forma de hacerlo. No es aconsejable saturar un informe con tablas y gráficos redundantes que, antes que claridad, crean confusión. Además la elección de determinada tabla o gráfico para mostrar los resultados, debe hacerse no sólo en función de las variables que relaciona, sino del lector a quien va dirigido el informe.

2.10 EL ANÁLISIS

La técnica estadística ofrece métodos y procedimientos objetivos que convierten las especulaciones de primera mano en aseveraciones cuya confiabilidad puede ser evaluada y ofrecer una premisa medible en la toma de una decisión.

Es el análisis donde se cristaliza la investigación. Esta es la fase de la determinación de los parámetros y estadísticos muestrales para las estimaciones e inferencias respecto a la población, el ajuste de modelos y las pruebas de las hipótesis planteadas, con el fin de establecer y redactar las conclusiones definitivas.

2.11 PUBLICACIÓN

Toda conclusión es digna de ser comunicada a un auditorio. Es más, hay otros estudiosos del mismo problema a quienes se les puede aportar información, conocimientos y otros puntos de vista acerca de él.

CUESTIONARIO

1. ¿Por qué se considera importante la estadística?

2. Enuncie las ramas en las que se divide la estadística y establezca su campo de acción.

3. Enumere las etapas del método estadístico.

4. ¿Por qué es importante la revisión bibliográfica en el desarrollo de una investigación estadística?.

5. ¿Qué es la hipótesis nula?.

6. Defina: Población, Muestra, Censo y Muestreo.

7. ¿Por qué usualmente se recurre al análisis a través de muestras y no de poblaciones?.

8. ¿Para qué se utiliza un muestreo piloto?.

9. ¿Con qué fin se critica una información?

10. ¿Cuáles son los componentes de una tabla?

3. Distribución de Frecuencias

Después de recoger toda la información correspondiente a la investigación, es decir, al agotar todo el trabajo de campo, nuestro escritorio se llena de un cúmulo de datos y cifras desordenadas los cuales, al ser tomados como observaciones individuales, dicen muy poco sobre la población estudiada; es, entonces, tarea del investigador "hacer hablar las cifras", comenzando por la clasificación y ordenación, consignando la información en tablas inteligibles que denominamos *distribuciones de frecuencias.*

3.1 DISTRIBUCIÓN DE FRECUENCIAS SIMPLE

Para una mayor sencillez, en la exposición del tema, nos valemos del siguiente ejemplo: Supongamos que en la fábrica de confecciones "La Hilacha", ha estallado un conflicto laboral y sus cincuenta operarias solicitan un aumento en el salario integral diario sopena de paralizar la fábrica.

El Gerente-propietario recoge la información respecto a la variable salario diario de sus 50 operarias y la relaciona en la tabla No 1.

Tabla No.1

SALARIO DIARIO DE 50 OPERARIAS EN LA FABRICA DE CONFECCIONES LA HILACHA (DATOS EN MILES DE PESOS)									
Obrera N°	Miles $/dia	Obrera N°	Miles $/dia	Obrera N°	Miles $/dia	Obrera N°	Miles $/dia	Obrera N°	Miles $/dia
1	52	11	54	21	55	31	56	41	52
2	54	12	51	22	55	32	53	42	57
3	55	13	54	23	52	33	57	43	56
4	54	14	55	24	55	34	54	44	51
5	53	15	54	25	53	35	53	45	58
6	56	16	56	26	57	36	50	46	55
7	54	17	52	27	54	37	55	47	53
8	58	18	54	28	55	38	52	48	54
9	51	19	53	29	53	39	53	49	53
10	54	20	55	30	55	40	54	50	56

Tabla No. 2

SALARIO DIARIO DE 50 OPERARIAS DE EN LA FABRICA DE CONFECCIONES "LA HILACHA" (DATOS EN MILES DE PESOS)				
Miles $/dia	Miles $/dia	Miles $/dia	Miles $/dia	Miles $/dia
52	54	55	56	52
54	51	55	53	57
55	54	52	57	56
54	55	55	54	51
53	54	53	53	58
56	56	57	50	55
54	52	54	55	53
58	54	55	52	54
51	53	53	53	53
54	55	55	54	56

Tabla No. 3

SALARIO DIARIO DE 50 OPERARIAS EN LA FABRICADE CONFECCIONES LA HILACHA (DATOS EN MILES DE PESOS)				
Miles $/dia	Miles $/dia	Miles $/dia	Miles $/dia	Miles $/dia
50	53	54	55	56
51	53	54	55	56
51	53	54	55	56
51	53	54	55	56
52	53	54	55	56
52	53	54	55	57
52	53	54	55	57
52	53	54	55	57
52	54	54	55	58
53	54	54	55	58

Tabla No. 4

DISTRIBUCION DE FRECUENCIAS DEL SALARIO DE 50 OPERARIAS		
MILES$/DIA	CONTEO	REPETICION
50	I	1
51	III	3
52	IIIII	5
53	IIIIIIII	9
54	IIIIIIIIIII	12
55	IIIIIIIII	10
56	IIIII	5
57	III	3
58	II	2
SUMA		50

Como se puede observar, hay una gran diferencia entre los datos brutos de la tabla No.1 y el ordenamiento y agrupamiento de la tabla No. 4.

Con el fin de obtener una mejor tabla interpretativa, introduciremos la siguiente simbología:

n: El tamaño de la muestra, es el número de observaciones.
X_i: La variable; es cada uno de los diferentes valores que se han observado.

La variable x_i, toma los x_1, x_2... x_m valores.

f_i: La frecuencia absoluta o simplemente frecuencia, es el número de veces que se repite la variable X_i; así f_1, es el número de veces que se repite la observación x_1, f_2 el número de veces que se repite la observación x_2 etc.

f_a: La frecuencia acumulada, se obtiene acumulando la frecuencia absoluta.

f_r: Frecuencia relativa; es el resultado de dividir c/u de las frecuencias absolutas por el tamaño de la muestra.

f_{ra}: Frecuencia relativa acumulada; se obtiene dividiendo la frecuencia acumulada entre el tamaño de la muestra.

Distribución Teórica de Frecuencias de n Observaciones

Variable X_i	frecuencia f_i	Frecuencia acumulada f_a	frecuencia relativa f_r	frecuencia relativa acumu f_{ra}
x_1	f_1	f_1	f_1/n	f_1/n
x_2	f_2	f_1+f_2	f_2/n	$(f_1+f_2)/n$
.	.	.	.	.
.				
.	.	.	.	.
.				
.	.	.	.	.
.				
x_i	f_i	$f_1+f_2+....+f_i$	f_i/n	$(f_1+f_2+....+f_i)/n$
.	.	.	.	.
.				
.	.	.	.	.
.				
x_m	f_m	$f_1+f_2+.....f_m$	f_m/n	$(f_1+f_2.....+f_m)/n$
	n		1.00	

Veamos el ejemplo que venimos trabajando:

Tabla No. 5

Distribución de Frecuencias del Salario Diario de 50 Obreras

Salario \$/dia X_i	Frecuencia f_i	Frecuencia acumulada fa	Frecuencia relativa fr	Frecuencia relati acumula fra
50	1	1	1/50= 0.02	1/50=0.02
51	3	4	3/50= 0.06	4/50=0.08
52	5	9	5/50= 0.10	9/50=0.18
53	9	18	9/50= 0.18	18/50=0.36
54	12	30	12/50=0.24	30/50=0.60
55	10	40	10/50=0.20	40/50=0.80
56	5	45	5/50= 0.10	45/50=0.90
57	3	48	3/50= 0.06	48/50=0.96
58	2	50	2/50= 0.04	50/50=1.00
Sumas	50		1.00	

En la práctica, cuando se tiene confianza en el ordenamiento, no son necesarias tantas tablas; se puede pasar de la tabla No1 directamente a la tabla No 6.

Tabla No. 6
Salario Diario de 50 Operarias de La Fabrica de Confecciones "La Hilacha"(Miles de Pesos)

$/DIA	X_i	f_i	fa	fr	fra
50		1	1	0.02	0.02
51		3	4	0.06	0.08
52		5	9	0.10	0.18
53		9	18	0.18	0.36
54		12	30	0.24	0.60
55		10	40	0.20	0.80
56		5	45	0.10	0.90
57		3	48	0.06	0.96
58		2	50	0.04	1.00
SUMAS		**50**		**1.00**	

Analizando las columnas porcentuales fr y fra se obtienen, entre otras las siguientes conclusiones:

- Sólo el 4% de las obreras gana el máximo salario/día de la fabrica, el cual corresponde a $58.000.00
- El salario diario mínimo ($50.000.00) lo gana únicamente una obrera, lo que constituye el 2% del personal asalariado.
- El 62% de las operarias tiene un salario diario entre $53.000.00 y $55.000.00
- El 60% de las obreras tiene un salario/día de $54.000.00 o menos.
- El 64% tiene un ingreso/día de $54.000.00 o más.

CUESTIONARIO Y EJERCICIOS PROPUESTOS

1. ¿Qué es frecuencia absoluta?

2. Cómo se obtiene:
2.1 ¿La frecuencia acumulada?

2.2 ¿La frecuencia relativa?
2.3 ¿La frecuencia relativa acumulada

3. En una distribución de frecuencias ¿se pueden establecer conclusiones porcentuales, utilizando solamente la frecuencia relativa? ¿Por qué?

4. La siguiente tabla relaciona las ausencias al trabajo de 50 obreras, durante el mes de octubre, en la fabrica de confecciones "la hilacha".

1	0	2	1	3	1	4	3	2	5
3	2	4	2	0	3	1	2	0	2
1	1	0	1	0	0	1	2	1	3
4	0	2	3	2	0	0	2	5	2
2	4	2	1	3	1	2	1	0	2

4.1 Construir una distribución de frecuencias simple.
4.2 Sacar 3 conclusiones.

5. Años de experiencia de las 50 operarias de la fabrica de confecciones "la hilacha"

4	6	5	6	4	6	5	5	6	5
5	5	8	8	8	6	9	6	5	7
7	9	3	2	7	4	5	7	7	3
6	7	7	7	8	3	6	6	7	6
4	6	8	5	6	6	7	5	7	4

Ordenar la Información y responder :
5.1 ¿Qué porcentaje de las obreras tiene experiencia inferior o igual a 6 años?
5.2 ¿Que porcentaje tiene experiencia entre 5 y 7 años (incluyendo los extremos)?

6.

Palabras por Minuto Escritas por un Grupo de Mecanógrafas

77	78	80	76	81	78	76	78	81	78	79	79	79
81	81	79	79	80	75	81	76	82	76	79	77	79
78	81	78	78	80	81	83	78	79	78	80	78	81
81	80	82	79	79	81	81	79	75	80	78	76	77
80	83	78	80	79	76	80	80	81	78	75	79	84
84	78	77	78	80	76	82	83	76	81	75	79	82
79	77	82	77	82	82	80	79	79	80	80	75	80
82	77	79	77	78	75	80	78	83	76	80	82	82
83	80	82	79	80	82	80	78	83	76	79	79	75
83	76	80	79	77	78	79	82	77	77	78	80	78
77	76	83	83	84	82	75	83	75	80	81	80	81
77	76	81	77	77	77	75	76	79	81	80	79	84
78	82	79	78	76	77	80	82	84	78	80	82	80
80	75	78	81	77	80	80	83	76	77	82	81	77
84	77	75	81	78	80	77	79	80	76	77	81	79
79	81	77	82	77								

Construir una distribución de frecuencias y resaltar 3 conclusiones

7. La siguiente tabla muestra, las respuestas obtenidas en un cuestionario aplicado a las obreras de la fábrica "La hilacha", respecto a la edad, estado civil, número de hijos, experiencia, años de estudio, ingresos diarios, gastos en educación y ausencias al trabajo en el último mes, así como una calificación del desempeño otorgada por el supervisor.

Obrera N°	Edad	Estado Civil	Numero de hijos	Expe-riencia	Esco-laridad	Miles $/dia	Gastos Educac.	ausen cias	Califi-cacion
1	24	soltera	2	4	5	52	5	3	1
2	24	soltera	2	5	5	54	6	2	1
3	27	casada	3	7	4	55	8	1	4
4	25	casada	3	6	4	54	9	1	3
5	24	viuda	1	5	3	53	3	2	2
6	28	soltera	0	7	8	56	1	1	4
7	29	u. Libre	1	5	3	54	2	2	3
8	35	soltera	0	9	9	58	0	0	5
9	30	casada	3	3	3	51	10	3	1
10	27	casada	3	6	3	54	9	2	2
11	28	soltera	1	7	6	54	3	2	3
12	25	u. Libre	2	3	3	51	6	5	1
13	27	soltera	0	6	7	54	1	1	2
14	30	soltera	0	7	7	55	1	1	3
15	27	soltera	0	6	5	54	2	2	3
16	36	soltera	0	8	8	56	3	1	4
17	26	soltera	1	4	3	52	2	3	2
18	29	viuda	2	6	4	54	5	2	2
19	26	soltera	2	5	4	53	5	3	2
20	28	soltera	0	7	9	55	4	2	3
21	28	soltera	0	7	8	55	4	1	3
22	31	soltera	1	7	6	55	4	2	3
23	22	soltera	2	4	3	52	7	3	1
24	25	u. Libre	1	7	6	55	3	1	3
25	25	viuda	3	5	3	53	7	2	2
26	40	soltera	0	8	9	57	3	1	5
27	39	casada	4	6	5	54	13	2	3
28	38	soltera	3	6	5	55	8	2	3
29	35	viuda	3	5	4	53	8	2	2
30	33	casada	3	7	4	55	9	0	3
31	33	soltera	1	8	6	56	4	0	4
32	32	soltera	2	5	4	53	6	2	2
33	32	soltera	0	9	8	57	2	0	4
34	31	soltera	1	6	5	54	3	1	3
35	30	casada	2	5	3	53	6	2	3
36	23	soltera	2	2	3	50	7	5	1
37	30	u. Libre	2	6	5	55	6	0	3
38	26	soltera	2	4	3	52	6	4	1
39	28	soltera	2	5	4	53	8	3	1
40	27	casada	2	6	4	54	8	1	2
41	26	u. Libre	3	4	3	52	11	4	1
42	38	soltera	1	8	9	57	3	0	4
43	38	soltera	0	7	8	56	5	0	4
44	31	viuda	2	3	3	51	6	4	1
45	39	soltera	1	8	9	58	3	0	4
46	36	casada	2	6	5	55	4	0	2
47	35	soltera	1	5	5	53	2	4	1
48	29	soltera	2	6	4	54	3	1	1
49	28	casada	2	6	5	53	7	3	1
50	29	soltera	1	7	6	56	3	0	3

Hacer las respectivas distribuciones de frecuencias, para cada una de las variables.

3. Distribución de Frecuencias

3.2 DISTRIBUCIÓN DE FRECUENCIAS POR INTERVALOS

Usualmente los valores de los datos no permiten un agrupamiento de ellos en una tabla de frecuencias simple, debido a que se encuentran distribuidos a través de todo el recorrido y el número de veces que se repite cada observación no es significativo en todos los casos, y en la mayoría de ellos su frecuencia es baja. Una tabla de frecuencias construida en estas condiciones, no presenta ninguna utilidad.

Ilustraremos el caso a través de un ejemplo, para ello, supongamos que la fabrica de baldosas"De las casas", con el objeto de ofrecer una garantía de su producto, desea hacer un estudio técnico de su producción, para lo cual extrae una muestra de 100 baldosas, cada una de las cuales se somete a una prueba de resistencia, destructiva cuyos datos expresados en Kg/ Cm^2, se relacionan a continuación:

Tabla No. 7
Resistencia en Kg/Cm^2 de 100 Baldosas de La Fabrica "De Las Casas"

478	458	683	780	736	448	591	555
339	694	478	498	310	537	592	549
666	239	398	720	648	533	586	321
313	644	495	122	521	368	531	472
415	291	621	253	763	746	323	575
210	480	223	433	444	437	360	559
425	459	418	351	361	183	383	259
419	655	487	135	370	345	282	578
425	436	634	450	223	479	161	337
420	422	282	439	449	321	452	444
391	569	460	308	477	463	367	251
487	610	470	469	392	517	359	527
540	504	542	369				

La clasificación en una distribución de frecuencias simple daría como resultante un ordenamiento de por lo menos 80 items; la mayoría de ellos con frecuencia unitaria.

Se hace necesario el agrupamiento en intervalos o clases que haga más compacta, manejable y presentable la información.

El número de clases y la amplitud de los intervalos los fija el investigador de acuerdo con el conocimiento que posea de la población, la necesidad de hacer comparación con otras investigaciones y la presentación de la información. Sin embargo, se recomienda que la información no sea demasiado compacta, lo cual le restaría precisión, ni demasiado dispersa, ya que no se tendría claridad.

En términos generales, es usual que el número de intervalos no sea inferior a 5 ni superior a 15. Struges propone que el número de clases o intervalos sea determinado por la expresión **m ≅1 + 3.3 log(n).***

La amplitud debe ser igual para todos los intervalos y, en lo posible, no se debe trabajar con clases abiertas.

3.3 REGLAS EMPÍRICAS PARA LA CONSTRUCCIÓN DE INTERVALOS

Cuando no se tiene experiencia en el manejo de la información es aconsejable seguir los pasos que se dan a continuación:

3.3.1 Determinar los datos de mayor y menor valor Xmax, Xmin.

3.3.2 Calcular el rango o recorrido $R = X\max - X\min$

3.3.3 Determinar el número de intervalos (m) y la amplitud de clase (A): $m \cong 1 + 3.3\log(n)$, Debe tenerse presente que m es un número natural. Luego se busca la amplitud A:

$$A > \frac{R}{m}$$,

3.3.4 Calcular el rango ampliado: $Ra = m.A$

3.3.5 Establecer la diferencia $a = Ra - R$, es decir la cantidad en que ha sido alterado el recorrido, la cual no debe ser superior a la amplitud.

(“a”) También puede ser definida como la cantidad positiva más pequeña que le hace falta al rango o recorrido para ser divisible exactamente por la amplitud

3.3.6 Distribuir adecuadamente la cantidad “a” de la siguiente manera:

Al valor X min se le resta aproximadamente $\frac{a}{2}$ y la parte restante se le suma a X max, obteniendo el límite inferior del primer intervalo y el límite superior del último, respectivamente.

$$X\min - \approx \frac{a}{2} = LIPI \text{ (Límite inferior del primer intervalo).}$$

$$X\max + \approx \frac{a}{2} = LSUI \text{ (Límite superior del último intervalo).}$$

3.3.7 Construir los intervalos, calcular los puntos medios o marcas de clase y hacer el agrupamiento de frecuencias.

Distribución Teórica de Frecuencias por Intervalos de n Observaciones

Inervalos*	Marca de clase. **X**	Frecuencia **fi**	Frecuencia acumulada. **fa**	Frecuencia relativa **fr**	Frecuencia relativa acumulada **fra**
LIPI-LIPI+A	X_1	f_1	f_1	f_1/n	f_1/n
LIPI+A-LIPI+2A	X_2	f_2	f_1+f_2	f_2/n	$(f_1+f_2)/n$
.	.	.	.		.
.	.	.	.		.
.		.			
.	.	.	.		.
.					
.	.	.	.		.
LIPI+ (m-1)A-LSUI	X_m	f_m	n	f_m/n	1.00
Sumas		n		1.00	.

n: Número de observaciones
LIPI: Límite inferior del primer intervalo

LSUI: Límite superior del último interval
X_i: Punto medio del intervalo, o marca de clase

* Con el fin de prever dobles conteos, quien clasifica deberá especificar si los intervalos son abiertos a la derecha o abiertos a la izquierda, en estas notas, trabajaremos con intervalos abiertos a la derecha; es decir, del tipo $a \leq X < b$, donde el límite superior no está incluido dentro de la clase.

Retomemos el ejercicio de la Tabla No. 7 y construyamos una distribución de frecuencia por intervalos.

3.3.1 $X\max = 780$,
$a < X \leq b$

3.3.2 Rango $R = X\max - X\min$,
$R = 780 - 122 = 658$

3.3.3 Número de intervalos $m \cong 1 + 3.3\log(n)$,
$m \cong 1 + 3.3\log(100)$,
$m \cong 1 + 3.3(2) = 7.6$

No es lógico tener 7.6 intervalos, por lo tanto se procede a aproximar el número de intervalos a un número natural cercano.

Aproximemos, $m = 7$, y busquemos la amplitud.

$$A > \frac{R}{m}, \quad A > \frac{658}{7}$$

Ya terminado el número de clases en m=7 encontramos que la amplitud debe ser mayor que 94. Fijémosla, entonces, en A = 100, que hace más manejable y presentable la tabla con la información.

3.3.4 Rango ampliado , $Ra = 7x100$.

3.3.5 Hemos alterado el rango original $R = 658$, cambiándolo por el rango ampliado

$Ra = 700$. La diferencia está representada por $a = Ra - R$ o sea $a = 700 - 658 = 42$

3.3.6 Tenemos por tanto, que distribuir adecuadamente la diferencia entre los rangos

$$X\min - \approx \frac{a}{2} = LIPI \quad ; 122 - 22 = 100 = LIPI$$

$$X\max + \approx \frac{a}{2} = LSUI \quad ; 780 + 20 = 800 = LSUI$$

Como se dijo antes, no estamos hablando de restar o sumar estrictamente $\frac{a}{2}$ sino una cantidad aproximada que brinde una buena presentación.

3.3.7 Construcción de los intervalos.

Tabla No. 8
Construcción de los Intervalos
para la Resistencia de las Baldosas

Intervalos	Marca de clase * X
100 – 200	150
200 – 300	250
300 – 400	350
400 – 500	450
500 – 600	550
600 – 700	650
700 – 800	750

Se puede desde luego, proceder a agrupar la información en los respectivos intervalos, haciendo la salvedad de que ninguno de los límites superiores de clase son considerados dentro de los intervalos.

Tabla No. 9
Distribución de Frecuencias por Intervalos
de la Resistencia de 100 Baldosas de la Fabrica “de las Casas”

Kg/Cm 2	X	f_i	fa	fr	fra
100 menos de 200	150	4	4	0.04	0.04
200 menos de 300	250	10	14	0.10	0.14
300 menos de 400	350	21	35	0.21	0.35
400 menos de 500	450	33	68	0.33	0.68
500 menos de 600	550	18	86	0.18	0.86
600 menos de 700	650	9	95	0.09	0.95
700 menos de 800	750	5	100	0.05	1.00
SUMAS		100		1.00	

Conclusiones:

- El 72% de las baldosas tiene una resistencia entre 300 y 600 Kg/Cm ^{2}.
- El 86% de las baldosas resiste menos de 600 Kg/Cm^2.
- Sólo el 5% resiste 700 o más Kg/Cm^2.

CUESTIONARIO Y EJERCICIOS PROPUESTOS

1. ¿Por qué se recurre al agrupamiento en distribuciones de frecuencias por intervalos?

2. ¿Cómo se determina el número de intervalos y la amplitud de ellos?

3. ¿Qué es una marca de clase?

4.

Consumo de agua, en m^{3}de 184 familias
n un barrio residencial de una ciudad durante el mes de octubre:

4	8	8	13	15	20	10	19	9	18	17
16	16	29	17	23	3	17	25	10	18	29
6	23	11	23	10	21	21	6	22	18	13
23	12	23	17	22	18	27	27	17	13	13
10	31	11	26	22	5	5	18	16	13	30
23	2	26	17	15	21	14	29	18	20	9
10	21	9	30	13	18	34	17	4	29	16
12	23	8	26	8	28	8	16	29	18	2
17	13	21	13	16	26	18	9	18	13	12
21	27	21	9	26	24	8	10	16	33	21
14	16	19	17	17	24	5	20	14	16	12
12	5	13	17	7	12	14	1	16	25	20
14	20	14	6	9	13	22	10	6	21	20
5	20	28	17	21	4	33	12	25	9	17
14	20	10	25	12	32	15	25	16	22	13
15	25	2	9	24	25	12	15	22	17	7
24	15	24	11	22	10	21	14			

Construir una distribución de frecuencias por intervalos.

4.1 Asumiendo el número de intervalos m = 8

4.2 Asumiendo el número de intervalos m = 9

4.3 Comparar las dos distribuciones y las conclusiones que de ellas se deriven.

5.

Calificaciones Obtenidas por 130 Estudiantes en un Examen de Estadística:

27	36	36	20	43	26	41	27	32	36	36	14	30
36	16	48	36	44	36	22	45	32	37	28	37	36
37	49	29	31	22	33	33	41	32	39	17	38	31
21	23	31	26	28	45	27	36	41	22	26	42	36
28	31	42	42	12	31	41	22	32	39	36	37	31
31	35	24	33	42	13	33	26	42	26	41	26	37
25	26	37	37	29	46	31	25	31	38	25	32	33
17	34	23	26	18	19	31	27	33	26	38	38	31
20	41	32	27	40	31	27	41	31	36	15	16	36
22	21	27	40	21	32	27	21	32	32	42	32	31

Construir una distribución de frecuencias por intervalos y resaltar cuatro (4) conclusiones.

---

* Ver Apéndice No. 2

4. Representación Gráfica

A pesar de la gran ayuda que prestan las tablas y cuadros con información organizada, no todos los públicos alcanzan a comprenderla o no disponen del tiempo suficiente para analizarla.

Es por ello que la mayoría de los investigadores acostumbran a reforzar la descripción a través de dibujos, generalmente con formas geométricas, que ayudan a visualizar el comportamiento de las variables tratadas.

4.1 DEFINICIÓN

Una gráfica o diagrama es un dibujo complementario a una tabla o cuadro, que permite observar las tendencias de un fenómeno en estudio y facilita el análisis estadístico de las variables allí relacionadas.

4.2 COMPONENTES DE UNA GRÁFICA

Una gráfica, al igual que un cuadro o una tabla, debe constar de:

4.2.1 Título adecuado: El cual debe ser claro y conciso, que responda a las preguntas: Qué relaciona, cuándo y dónde se hicieron las observaciones.

4.2.2 El cuerpo: o gráfico en sí, cuya elección debe considerar el o los tipos variables a relacionar, el público a quien va dirigido y el diseño artístico del gráfico.

4.2.3 Notas de pie de gráfico: Donde se presentan aclaraciones respecto al gráfico, las escalas de los ejes, o se otorgan los créditos a las fuentes respectivas.

Es de anotar que por medio de gráficos tendenciosos se pueden deformar o resaltar situaciones o estados, que presentados en un gráfico apropiado, mostrarían un comportamiento normal.

Generalmente una información es distorsionada por algunas de las siguientes causas:

4.2.1.1 La relación entre los ejes no es la mas apropiada (ver gráficos No.1 y No.2.

4.2.1.2 Gráficos con escalas desproporcionadas, o mala elección del punto de origen (ver gráfico No.3).

Variación de La Inflación en Colombia 1995-2000

1995	1996	1997	1998	1999	2000
19.46	21.63	17.68	16.7	9.23	7.81

Gráfico No. 1

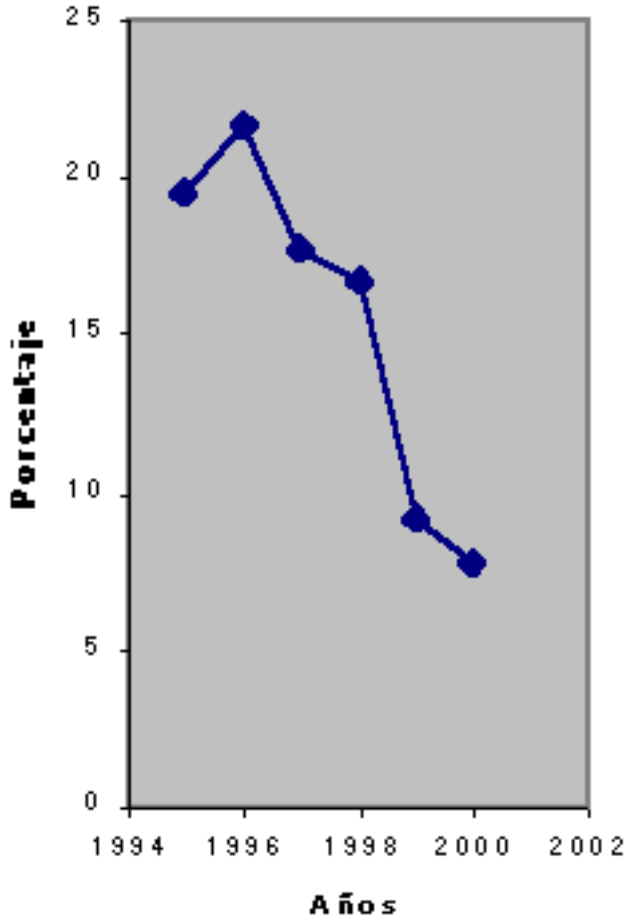


Gráfico No. 2

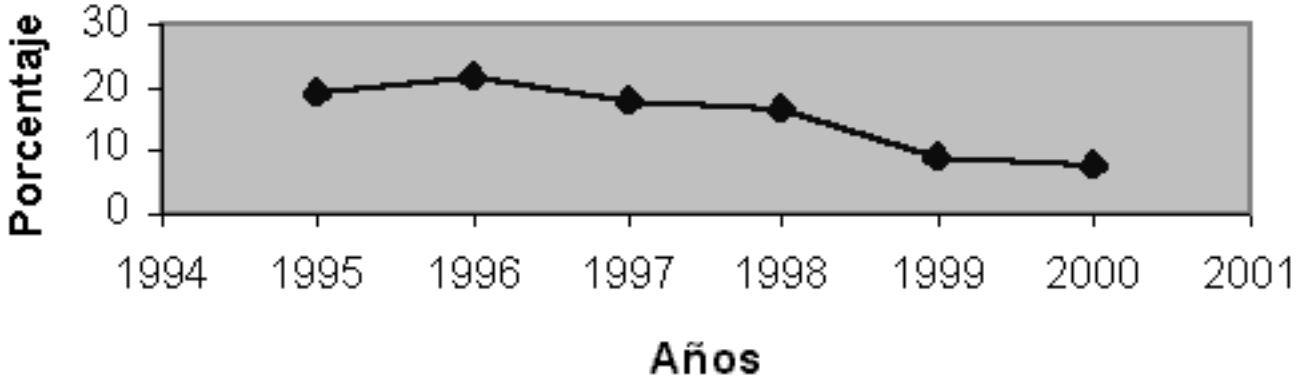
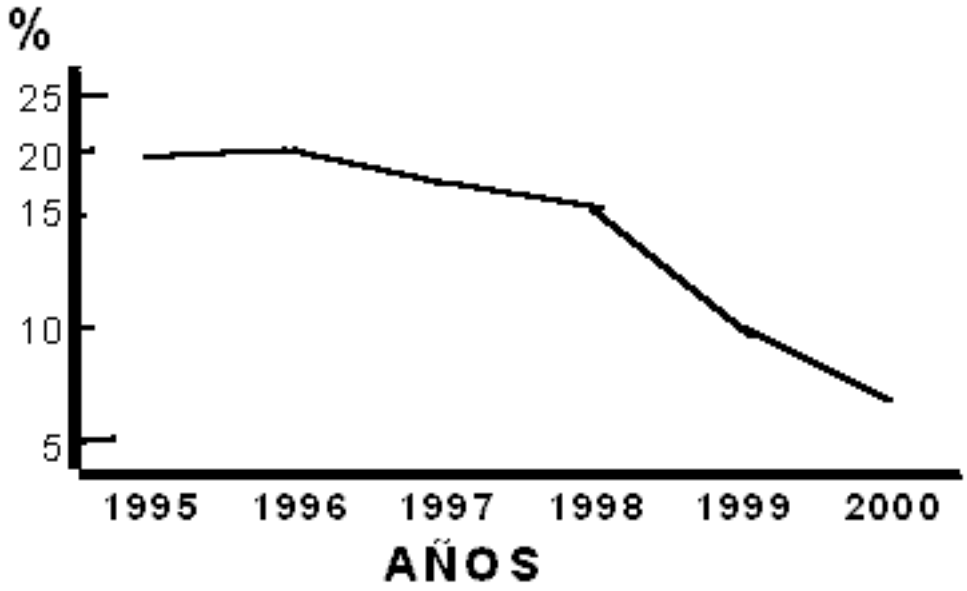


Gráfico No. 3



Como se puede observar, el gráfico No.1 “realza” el decrecimiento de la variable inflación, mientras que el No.2 intenta mostrar una estabilización o decrecimiento parsimonioso.

Los dos dibujos son incorrectos debido a que no conservan una proporción adecuada entre sus ejes. Sin embargo, el gráfico No. 3 tiene una buena proporción entre los ejes. Pero, la distorsión se debe a la mala numeración en el eje “Y” pues, el punto de origen O ha sido eliminado y asignado un valor arbitrario, la escala es inadecuada para resaltar el decrecimiento inflacionario de los dos últimos periodos.

Ambas situaciones son erróneas o tendenciosas y se deben corregir asignando escalas apropiadas a los ejes y utilizando la siguiente regla:

$$\frac{Lx}{Ly} = \frac{4}{3} \qquad Ly = \frac{3}{4} Lx$$

Donde: Lx: Longitud del eje horizontal

Ly: Longitud del eje vertical

“La longitud del eje vertical es igual a tres cuartos de la longitud del eje horizontal”.

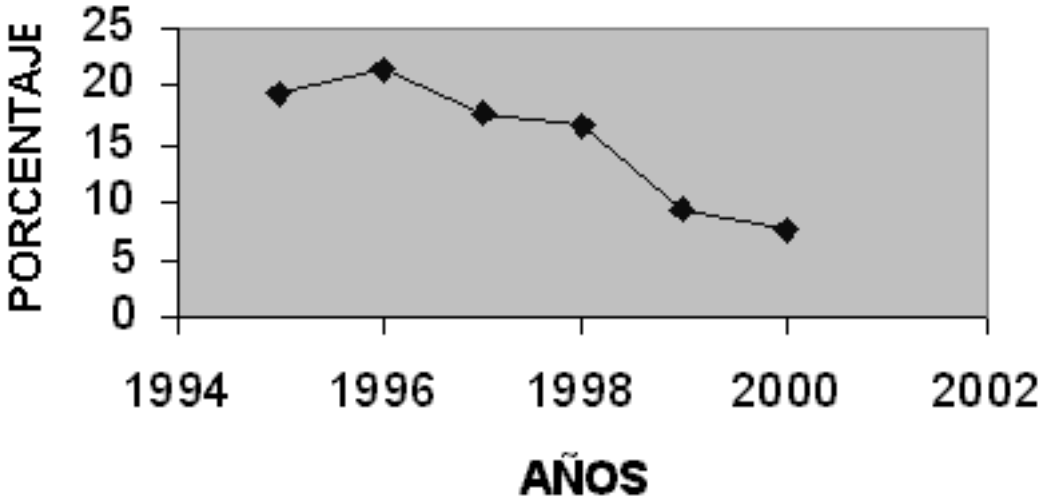
4.3 PRINCIPALES TIPOS DE GRÁFICOS

Existe una gran cantidad de gráficos para la representación de datos estadísticos, ya que de ellos depende el diseño artístico de quien los elabora, así como de su imaginación al combinar varios tipos de ellos, como forma de presentar una información.

Entre los gráficos más comunes tenemos:

4.3.1 Gráfico de Líneas: Usado básicamente para mostrar el comportamiento de una variable cuantitativa a través del tiempo. El gráfico de líneas consiste en segmentos rectilíneos unidos entre sí, los cuales resaltan las variaciones de la variable por unidad de tiempo. Para su construcción ha de procederse de la siguiente manera: en el eje de las ordenadas se marcan los puntos de acuerdo con la escala que se esté utilizando. En el caso de una escala aritmética, distancias iguales en el eje, representan distancias iguales en la variable.

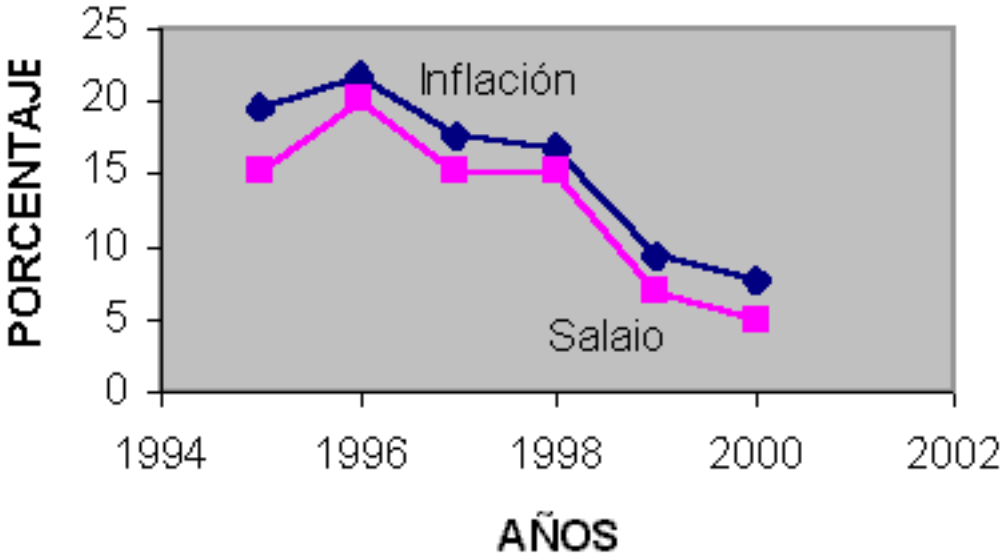
Variación de la Inflación en Colombia
1995 -2000



El eje de la variable X se divide en unidades de tiempo iguales, teniendo presente el número de ítems que ha de presentarse, así como la longitud del eje. Es de anotar la conveniencia de mostrar la interrupción y acercamiento del eje a su origen cuando esto haya ocurrido.

4.3.2 Gráfico de Líneas Compuesto: Cuando se tienen varias variables a representar, con el fin de establecer comparaciones entre ellas (siempre que su unidad de medida sea la misma); se utiliza plasmarlos en un sólo gráfico, el cual es el resultado de representar varias variables en un mismo plano.

Variación de la Inflación y el Salario de la Hilacha



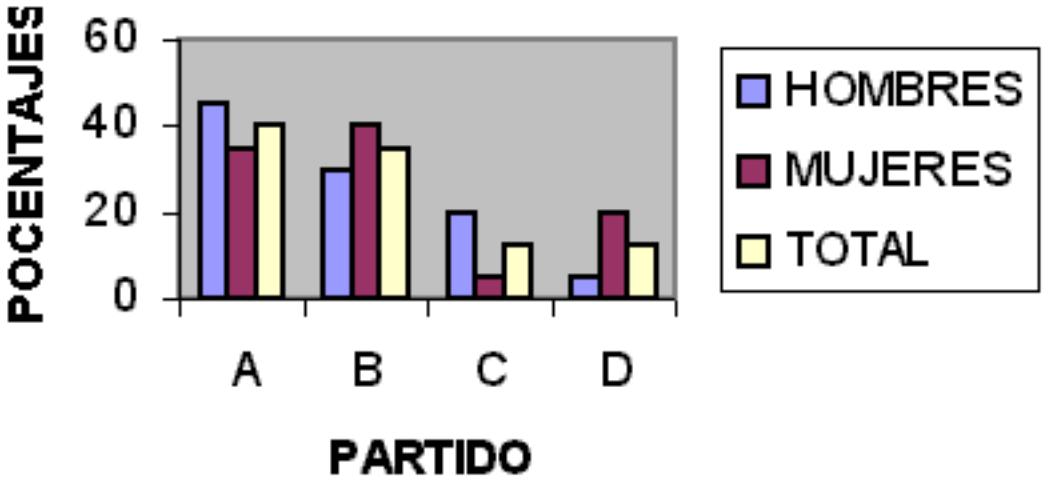
4.3.3 Gráfico de Barras: El gráfico de barras, como su nombre lo indica, está constituido por barras rectangulares de igual ancho, conservando la misma distancia de separación entre sí. Se utiliza básicamente para mostrar y comparar frecuencias de variables cualitativas o comportamientos en el tiempo, cuando el número de ítems es reducido.

Número de Hijos de 50 Obreras en "La Hilacha"

Éstos gráficos suelen ser de barras verticales, aunque se pueden utilizar de forma horizontal.

4.3.4 Gráfico de Barras Compuesto

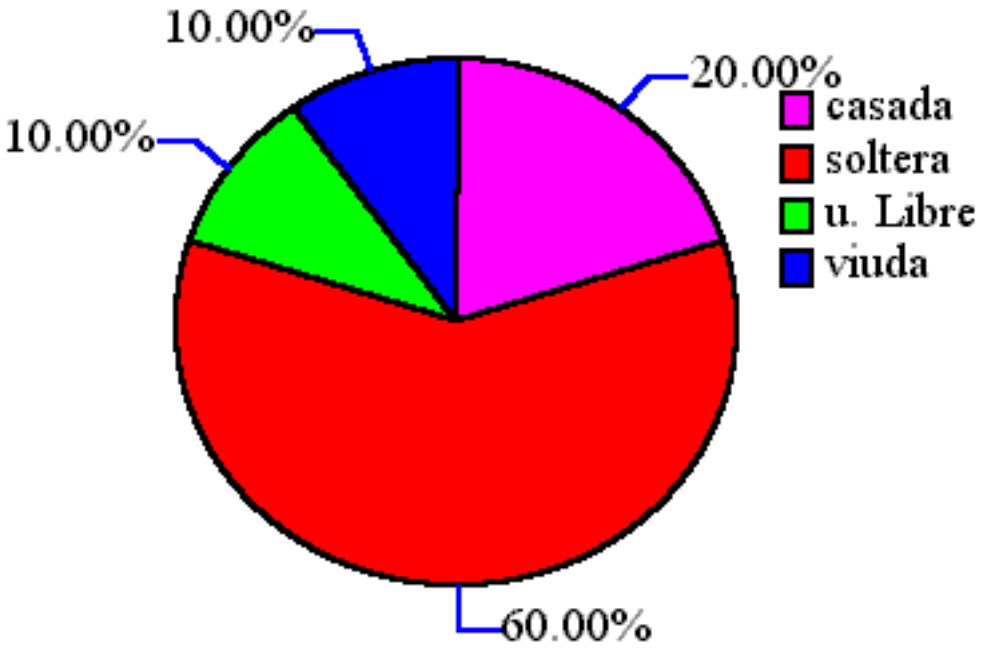
Preferencias de Partido Según Sexo



4.3.5 Gráfico de Sectores Circulares Usualmente llamado gráfico de pastel, debido a su forma característica de una circunferencia dividida en cascos, por medio de radios que dan la sensación de un pastel tajado en porciones.

Se usa para representar variables cualitativas en porcentajes o cifras absolutas cuando el número de ítems no es superior a 5 y se quiere resaltar uno de ellos. Para su construcción se procede de la siguiente forma: La circunferencia tiene en su interior 360 grados, los cuales hacemos corresponder al total de la información, es decir al 100%; luego, para determinar el número de grados correspondiente a cada componente se multiplica el porcentaje respectivo por 360 y se divide por 100, los cuales se miden con la ayuda de un transportador para formar los casquetes de los diferentes ítems.

Estado Civil de 50 Operarias de "La Hilacha"

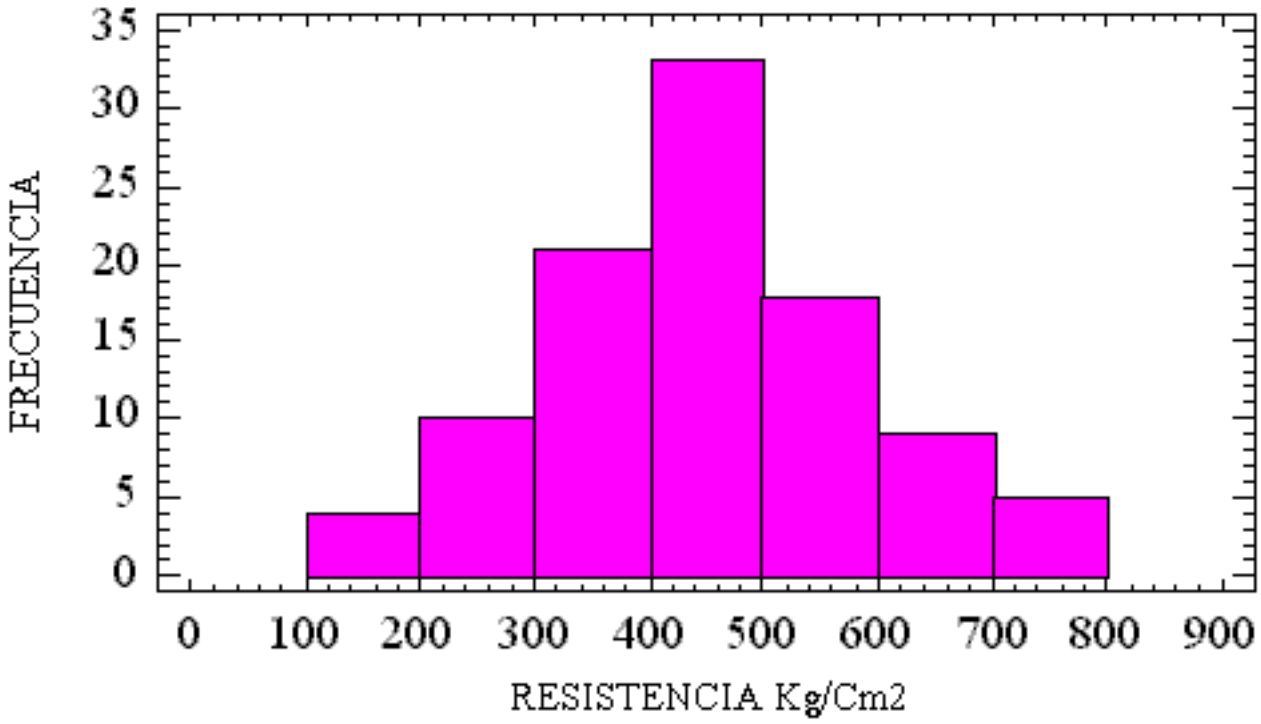


4.3.6 Histograma de Frecuencias: Para la construcción de un histograma de frecuencias de

fácil interpretación y que no falsee la información, debe disponerse de una distribución de frecuencias por intervalos con amplitud igual para cada clase o intervalo. En el eje de las abscisas procedemos a representar los intervalos de la variable, y en el eje de las ordenadas las frecuencias de cada clase.

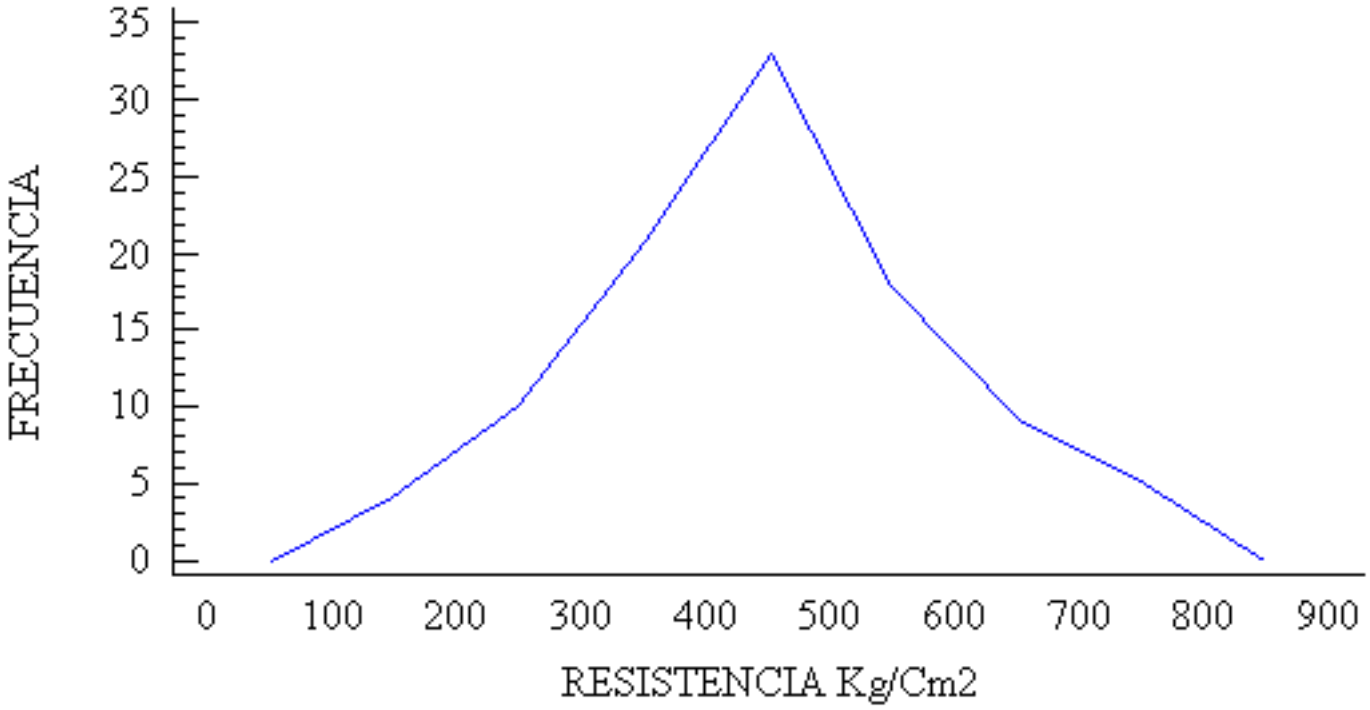
El histograma se construye dibujando barras contiguas que tienen como base la amplitud de cada intervalo y como alturas las frecuencias respectivas.

Histograma de Frecuencias de la Resistencia de 100 Baldosas



4.3.7 Polígono de Frecuencias

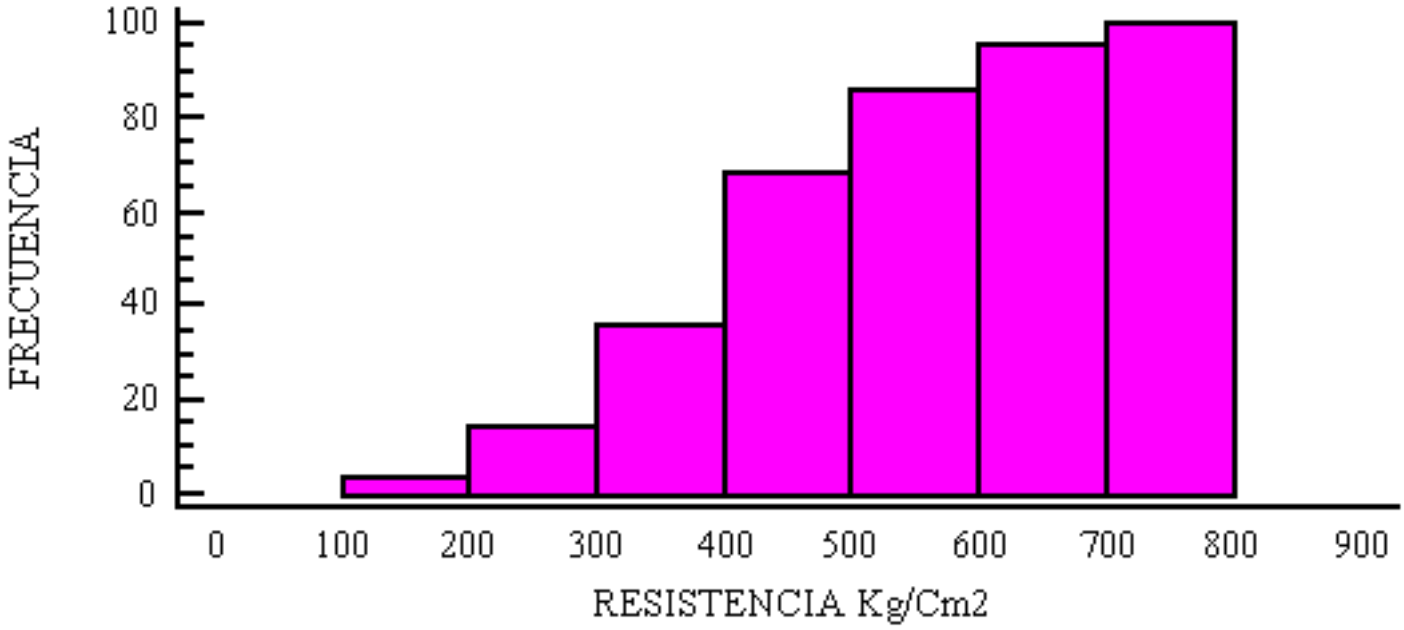
Resistencia de 100 Baldosas



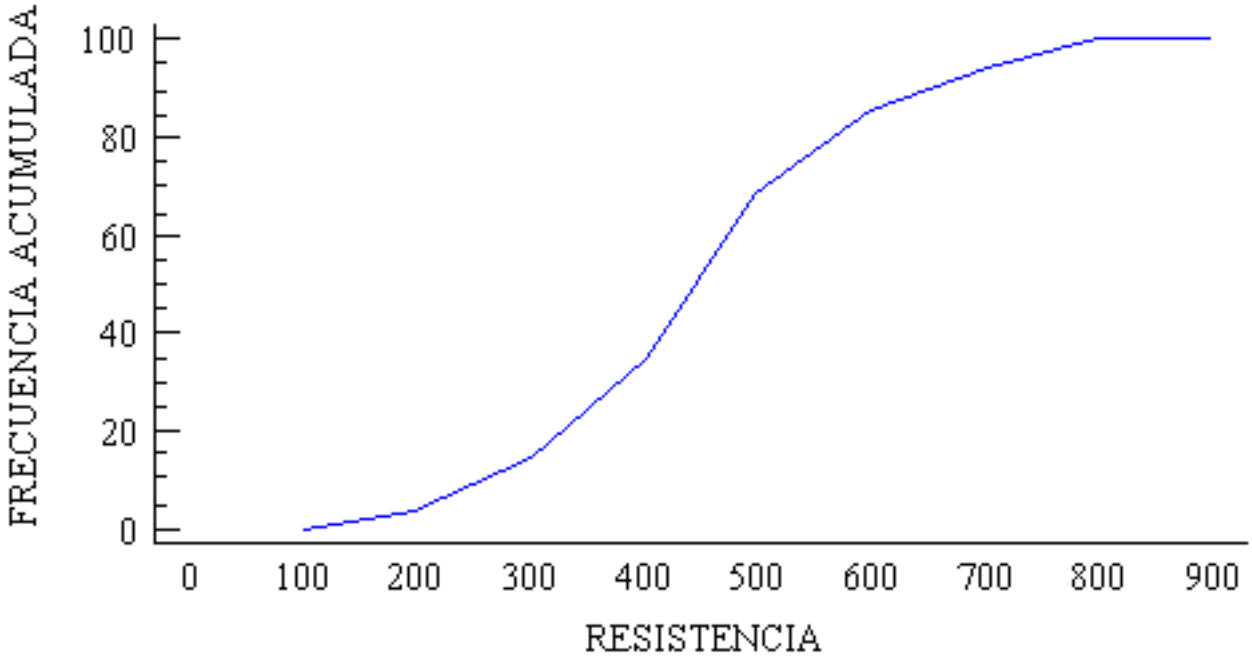
Para la construcción de un polígono de frecuencias, se marcan los puntos medios de cada uno los intervalos en la parte superior de cada barra del histograma de frecuencias, los cuales se unen con segmentos de recta.

4.3.8 Histograma de Frecuencias Acumuladas El histograma de frecuencias acumuladas también es obtenido a partir de una distribución de frecuencias, tomando en el eje horizontal las clases de la variable, y en el eje vertical las frecuencias acumuladas correspondientes a cada intervalo

Resistencia de 100 Baldosas



Resistencia de 100 Baldosas



CUESTIONARIO Y EJERCICIOS PROPUESTOS

1. ¿Cuál es el objetivo de un gráfico?

2. Describa los componentes de una gráfica .

3. ¿Cuáles son las principales causas de distorsión de la información de un gráfico?

4. ¿ Cuál debe de ser la proporción entre los ejes del plano cartesiano para la construcción de un gráfico?

5. Para los ejercicios 4 y 5 del capitulo 3, numeral 3.2 construir:
- **5.1** Un histograma de frecuencias
- **5.2** Un polígono de frecuencias.
- **5.3** Un histograma de frecuencias acumuladas
- **5.4** Un polígono de frecuencias acumuladas

6.

Costo Promedio del Consumo de Energía
de la Fábrica de Confecciones "La Hilacha".

Año	Costo mes
1995	30000
1996	80000
1997	70000
1998	130000
1999	140000
2000	140000
2001	160000

Construir un gráfico de líneas para esta información.

7.

Índice de Precios al Consumidor 1999-2001

Año	MES	Valor Índice	Variación Mes	Variación Acumulada	Variación 12 meses
1999	1	102.21	2.209	2.21	17.18
1999	2	103.94	1.698	3.94	15.38
1999	3	104.92	0.938	4.92	13.51
1999	4	105.74	0.784	5.74	11.17
1999	5	106.25	0.479	6.25	9.98
1999	6	106.55	0.279	6.55	8.96
1999	7	106.88	0.312	6.88	8.78
1999	8	107.41	0.495	7.41	9.28
1999	9	107.76	0.331	7.76	9.33
1999	10	108.14	0.350	8.14	9.32
1999	11	108.66	0.479	8.66	9.65
1999	12	109.23	0.529	9.23	9.23
2000	1	110.64	1.289	1.29	8.25
2000	2	113.19	2.302	3.62	8.89
2000	3	115.12	1.711	5.39	9.73
2000	4	116.27	0.996	6.44	9.96
2000	5	116.88	0.521	7.00	10.00
2000	6	116.85	-0.019	6.98	9.68
2000	7	116.81	-0.039	6.94	9.29
2000	8	117.18	0.316	7.27	9.10
2000	9	117.68	0.426	7.73	9.20
2000	10	117.86	0.153	7.90	8.99
2000	11	118.24	0.328	8.25	8.82
2000	12	118.79	0.460	8.75	8.75
2001	1	120.04	1.051	1.05	8.49
2001	2	122.31	1.893	2.96	8.06
2001	3	124.12	1.481	4.49	7.81
2001	4	125.54	1.148	5.69	7.98
2001	5	126.07	0.418	6.13	7.87
2001	6	126.12	0.040	6.17	7.93

Graficar : El valor del índice, la variación mensual y la variación anual, en función del tiempo.

8. Resultados electorales en Colombia, en la elección de presidente de la república para el período 1986-1990:

Liberal 58%
Conservador 36%
Otros 6%

Construir un gráfico apropiado para esta información.

5. Medidas de Tendencia Central

En los capítulos anteriores, nos referimos a la clasificación, ordenación y presentación de datos estadísticos, limitando el análisis de la información a la interpretación porcentual de las distribuciones de frecuencia.

El análisis estadístico propiamente dicho, parte de la búsqueda de parámetros sobre los cuales pueda recaer la representación de toda la información.

Las medidas de tendencia central, llamadas así porque tienden a localizarse en el centro de la información, son de gran importancia en el manejo de las técnicas estadísticas, sin embargo, su interpretación no debe hacerse aisladamente de las medidas de dispersión, ya que la representabilidad de ellas está asociada con el grado de concentración de la información.

Las principales medidas de tendencia central son:

5.1 Media aritmética.
5.2 Mediana
5.3 Moda.

5.1 MEDIA ARITMÉTICA

Cotidiana e inconscientemente estamos utilizando la media aritmética. Cuando por ejemplo, decimos que un determinado fumador consume una cajetilla de cigarrillos diaria, no aseguramos que diariamente deba consumir exactamente los 20 cigarrillos que contiene un paquete sino que es el resultado de la observación, es decir, dicho sujeto puede consumir 18, un día; 19 otro; 20, 21, 22; pero según nuestro criterio, el número de unidades estará alrededor de 20.

Matemáticamente, la media aritmética se define como la suma de los valores observados dividida entre el número de observaciones.

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_i + \ldots + x_n}{n} = \frac{\sum_{1}^{n} x_i}{n}$$

$\overline{X}$: Media aritmética de la variable X
x_i: Valores de la variable X
n: Número de observaciones

$\sum$: Signo de sumatoria, indica que se debe sumar

Ejemplo: Cantidad de cigarrillos consumidos por un fumador en una semana.

Lunes: 18
Martes: 21
Miércoles: 22
Jueves: 21
Viernes: 20
Sábado: 19
Domingo: 19

Entonces la media aritmética es.

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_i + \ldots + x_n}{n} = \frac{\sum_1^n x_i}{n}$$

$$\overline{X} = \frac{\sum_1^7 x_i}{7}$$

$$\overline{X} = \frac{18 + 21 + 22 + 21 + 20 + 19 + 19}{7} = 20$$

El fumador consume en promedio 20 cigarrillos diarios.

Cuando la variable está agrupada en una distribución de frecuencias, la media aritmética se calcula por la fórmula:

$$\overline{X} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_i f_i + \ldots + x_m f_m}{n} = \frac{\sum_1^m x_i f_i}{n}$$

Ejemplo:

Cantidad de Cigarrillos Consumidos por un Fumador en una Semana Dada:

Cantidad Xi	Frecuencia fi
18	1
19	2
20	1
21	2
22	1
	7

$$\overline{X} = \frac{\sum_{1}^{m} x_i f_i}{n} = \frac{18(1) + 19(2) + 20(1) + 21(2) + 22(2)}{7} = 20$$

$$\overline{X} = 20 \, cigarrillos/dia$$

Ejemplo:

Calculo de La Media Aritmética.
El Salario/día de 50 Operarias

Miles $/día Xi	fi	Xifi
50	1	50
51	3	153
52	5	260
53	9	477
54	12	648
55	10	550
56	5	280
57	3	171
58	2	116
Sumas	50	2705

$$\overline{X} = \frac{\sum_{1}^{9} x_i f_i}{50} \quad ,$$

$$\overline{X} = \frac{2705}{50} = 54.1 \ ,$$

= 54.100 pesos/día

Si la información está relacionada en una distribución de frecuencias por intervalos, se toman como valores de la variable las marcas de clase de los intervalos, entiéndase por marca de clase el punto medio entre los límites de cada clase o intervalo.

Ejemplo:

Cálculo de La Media Aritmética de la Resistencia de 100 Baldosas

Resistencia Kg/cm2	X	f_i	$x_i f_i$
100 y menos de 200	150	4	600
200 y menos de 300	250	10	2500
300 y menos de 400	350	21	7350
400 y menos de 500	450	33	14850
500 y menos de 600	550	18	9900
600 y menos de 700	650	9	5850
700 y menos de 800	750	5	3750
Sumas		100	44800

$$\overline{X} = \frac{\sum_{1}^{7} x_i f_i}{100} = \frac{44800}{100} = 448$$

La resistencia promedio de las 100 baldosas es de 448 Kg/Cm².

5.1.1 Propiedades de la Media Aritmética

5.1.1.1 La suma de las diferencias de los datos con respecto a la media aritmética es igual cero.

$$\sum_{1}^{n} (x_i - \overline{X}) = 0$$

Demostración:

$$\sum_{1}^{n} (x_i - \overline{X}) = \sum_{1}^{n} x_i - \sum_{1}^{n} \overline{X}$$

pero

$$\overline{X} = \frac{\sum_{1}^{n} x_i}{n}$$

$$n\overline{X} = \sum_{1}^{n} x_i$$

Como

$$\sum_{1}^{n} \overline{X} = n\overline{X}$$

$$\sum_{1}^{n} (x_i - \overline{X}) = n\overline{X} - n\overline{X} = 0$$

$$\sum_{1}^{n} (x_i - \overline{X}) = 0$$

Ejemplo de Comprobación:

En el Ejercicio del Fumador Cuya Media Aritmética es de 20 Cigarrillos / día:

X	$x - \overline{X}$	
18	18-20 =	-2
21	21-20 =	1
22	22.-20 =	2
21	21-20 =	1
20	20-20 =	0
19	19-20 =	-1
19	19-20 =	-1
Sumas		0

Para una distribución de frecuencias:

Salario/día de 50 Operarias en La Fábrica de Confecciones "La Hilacha" (Miles De Pesos)

Miles $ X_i	f_i	$(x_i - \overline{X})$	$(x_i - \overline{X}) f_i$
50	1	50-54.1= -4.1	-4.1
51	3	51-54.1= -3.1	-9.3
52	5	52-54.1= -2.1	-10.5
53	9	53-54.1= -1.1	-9.9
54	12	54-54.1= -0.1	-1.2
55	10	55-54.1= 0.9	9.0
56	5	56-54.1= 1.9	9.5
57	3	57-54.1= 2.9	8.7
58	2	58-54.1= 3.9	7.8
Sumas	50		0

5.1.1.2 La suma de las diferencias cuadráticas de los datos, con respecto a la Media Aritmética, es mínima.

Quiere decir esta propiedad que cualquier otro parámetro p, diferente a la media aritmética hace mayor la expresión:

$$\sum_1^n (x_i - p)^2 \quad \text{que} \quad \sum_1^n (x_i - \overline{X})^2$$

Para
$\overline{X} \neq p$

Demostración:

Debemos, entonces, probar que:

$$\sum_1^n (x_i - p)^2 > \sum_1^n (x_i - \overline{X})^2$$

veamos:

$$(x_i - p) = (x_i - p) + (\overline{X} - \overline{X})$$

$$(x_i - p) = (x_i - \overline{X}) + (\overline{X} - p)$$

$$(x_i - p)^2 = [\,(x_i - \overline{X}) + (\overline{X} - p)\,]^2$$

$$(x_i - p)^2 = (x_i - \overline{X})^2 + 2(x_i - \overline{X})(\overline{X} - p) + (\overline{X} - p)^2$$

$$\sum_1^n (x_i - p)^2 = \sum_1^n \left[x_i - \overline{X})^2 + 2(x_i - \overline{X})(\overline{X} - p) + (\overline{X} - p)^2\right]$$

$$\sum_1^n (x_i - p)^2 = \sum_1^n (x_i - \overline{X})^2 + \sum_1^n 2(x_i - \overline{X})(\overline{X} - p) + \sum_1^n (\overline{X} - p)^2$$

$$\sum_1^n (x_i - p)^2 = \sum_1^n (x_i - \overline{X})^2 + 2(\overline{X} - P)\sum_1^n (x_i - \overline{X}) + \sum_1^n (\overline{X} - p)^2$$

pero $\sum_1^n (x_i - \overline{X}) = 0$ (propiedad a.)

entonces:

$$\sum_1^n (x_i - p)^2 = \sum_1^n (x_i - \overline{X})^2 + \sum_1^n (\overline{X} - p)^2$$

$$\sum_1^n (x_i - p)^2 = \sum_1^n (x_i - \overline{X})^2 + n(\overline{X} - p)^2$$

como

$\overline{X} \neq p \qquad (\overline{X} - p)^2 > 0 \qquad n(x_i - p)^2 > 0$

luego

$$\sum_1^n (x_i - p)^2 > \sum_1^n (x_i - \overline{X})^2$$

5.1.1.3 Si a cada uno de los resultados le sumamos o le restamos una constante C , la Media Aritmética queda alterada en esa constante.

Demostración:

Tenemos los datos $x_1, x_2, \ldots \ldots x_n$. Cuya media aritmética es $\overline{X}$

sea

$$y_1 = (x_1 \pm c),\ y_2 = (x_2 \pm c),\ y_i = (x_{ni} \pm c), \ldots, y_n = (x_n \pm c),$$

La media aritmética de la nueva variable es:

$$\overline{Y} = \frac{\sum_1^n y_i}{n}$$

$$\overline{Y} = \frac{\sum_1^n y_i}{n} = \frac{\sum_1^n (x_i \pm C)}{n} = \frac{\sum_1^n x_i}{n} + \frac{\sum_1^n C}{n} = \frac{\sum_1^n x_i}{n} + \frac{nC}{n}$$

entonces

$$\overline{Y} = \overline{X} \pm C$$

En el ejemplo de las baldosas, $\overline{X} = 448$, a cada uno de los datos restémosle una constante $C = 450$.

$$Y_i = X_i - C$$

Resistencia Kg/cm2	X	f_i	$Yi=X-450$	$Y_i f_i$
100 y menos de 200	150	4	-300	-1200
200 y menos de 300	250	10	-200	-2000
300 y menos de 400	350	21	-100	-2100
400 y menos de 500	450	33	0	0
500 y menos de 600	550	18	100	1800
600 y menos de 700	650	9	200	1800
700 y menos de 800	750	5	300	1500
Sumas		100		-200

$$\overline{Y} = \frac{-200}{100} = -2 \ , \ \overline{Y} = \overline{X} \pm C \ \rightarrow \ \overline{X} = \overline{Y} \mp C$$

$$\overline{X} = -2 + 450 = 448 kg/cm^2$$

5.1.1.4 Si cada uno de los datos se multiplica por una constante k, entonces la media aritmética queda multiplicada por esa constante:

Tenemos los datos $x_1, x_2, \ldots \ldots x_n$ cuya media aritmética es $\overline{X}$

sea

$$y_1 = kx_1 \ , \quad y_2 = kx_2 , \quad y_i = kx_i \ \ldots \ , \ y_n = kx_n ,$$

$$\overline{Y} = \frac{\sum_1^n y_i}{n} = \frac{\sum_1^n kx_i}{n} = \frac{k\sum_1^n x_i}{n} = k\overline{X} \ \rightarrow \ \overline{X} = \frac{\overline{Y}}{k}$$

Si multiplicamos cada una de las resistencias de las 100 baldosas por

una constante $k = \frac{1}{100}$ tenemos:

Resistencia Kg/cm2	X	f_i	Yi=(1/100)Xi	$Y_i f_i$
100 y menos de 200	150	4	1.5	6
200 y menos de 300	250	10	2.5	25
300 y menos de 400	350	21	3.5	73.5
400 y menos de 500	450	33	4.5	148.5
500 y menos de 600	550	18	5.5	99
600 y menos de 700	650	9	6.5	58.5
700 y menos de 800	750	5	7.5	37.5
Sumas		100		448

$$\overline{Y} = \frac{448}{100} = 4.48 \rightarrow \overline{X} = 100\overline{Y} \rightarrow \overline{X} = 100(4.48) = 448 kg/cm^2$$

5.1.2 Media Aritmética con Cambio de Origen y de Escala

En estadística es usual la transformación de variables utilizando las dos últimas propiedades:

C = un valor de tendencia central (media, mediana, moda o cualquier otro parámetro.

k = generalmente la desviación standar, desviación media, la amplitud etc.

Sea

$Y_i = \frac{1}{k}(X_i - C)$ para nuestro ejemplo C = 450, k = 100

Resistencia Kg/cm2	X	f_i	Yi=(X-450)/100	$Y_i f_i$
100 y menos de 200	150	4	-3	-12
200 y menos de 300	250	10	-2	-20
300 y menos de 400	350	21	-1	-21
400 y menos de 500	450	33	0	0
500 y menos de 600	550	18	1	18
600 y menos de 700	650	9	2	18
700 y menos de 800	750	5	3	15
Sumas		100		-2

A la nueva variable “Y” le calculamos la media aritmética.

$$\overline{Y} = \frac{\sum_{1}^{m} y_i f_i}{n} = \frac{-2}{100} = -0.02 \qquad \overline{X} = k\overline{Y} + C$$

$$\overline{X} = 100(-.02) + 450 = 448 kg/cm^2$$

5.1.3 Media Aritmética Ponderada

Hemos visto que la Media Aritmética se calcula con base a la magnitud de los datos, otorgándoles igual importancia a cada uno de ellos. Sin embargo en muchas ocasiones la magnitud del dato esta ponderada con un determinado peso que lo afecta relativamente.

La Media Aritmética ponderada tiene en cuenta la importancia relativa de cada uno de los datos, para lo cual la definimos con la siguiente expresión:

$$\overline{X}_w = \frac{\sum_{1}^{n} x_i w_i}{\sum_{1}^{n} w_i}$$

donde

$\overline{X}_w$: Media aritmética ponderada

x_i: Valor de la variable X

w_i: Ponderación del ítem x_i

Ejemplo:

Las calificaciones de un estudiante están conformadas por los siguientes factores:

Un examen cuyo valor es 40% en el cual obtuvo una nota de 4.5, un trabajo de consulta con ponderación del 10% y calificación de 1.0, una exposición equivalente al 15% con nota de 2.0, y por último una investigación con valor del 35% calificada con 3.5.

$$\overline{X}_w = \frac{\sum_1^n x_i w_i}{\sum_1^n w_i}$$

entonces la nota definitiva es:

5. Medidas de Tendencia Central

5.2 LA MEDIANA

Otra medida de tendencia central, utilizada principalmente en estadística no paramétrica, es la mediana, la cual no se basa en la magnitud de los datos, como la media aritmética, sino en la posición central que ocupa en el orden de su magnitud, dividiendo la información en dos partes iguales, dejando igual número de datos por encima y por debajo de ella.

5.2.1 La Mediana Cuando los datos no están Agrupados en Intervalos.

Partiendo de la información bruta, ordenamos los datos ascendente o descendentemente:

$x_1, x_2, x_3, \ldots\ldots\ldots\ldots, x_i, \ldots\ldots\ldots\ldots\ldots, x_n$ se define

Mediana = $Me = x_{(\frac{n+1}{2})}$, si n es impar ó

Mediana = $Me = \dfrac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$, si n es par

En el ejercicio de los cigarrillos, consumidos por un fumador tenemos lunes 18, martes 21, miércoles 22, jueves 21, viernes 20, sábado 19, y domingo 19. Ordenando ascendentemente :

$x_1 = 18, x_2 = 19, x_3 = 19, x_4 = 20, x_5 = 21, x_6 = 21, x_7 = 22$

n, es impar, entonces

$Me = x_{(\frac{n+1}{2})} = x_{(\frac{7+1}{2})} = x_4 = 20$

Veamos cuando n es par:

Consumo mensual de agua, en m^3, por la fábrica de confecciones “la hilacha”.

Enero=	10,	Mayo=	14,	Septiembre=	18,
Febrero=	12,	Junio=	19,	Octubre=	22,
Marzo=	15,	Julio=	17,	Noviembre=	15,
Abril=	18,	Agosto=	18,	Diciembre=	13

$$x_1 = 10\ ,\ x_2 = 12\ ,\ x_3 = 13\ ,\ x_4 = 14\ ,\ x_5 = 15$$
$$x_6 = 15\ ,\ x_7 = 17\ ,\ x_8 = 18\ ,\ x_9 = 18\ ,\ x_{10} = 18$$
$$x_{11} = 19\ ,\ x_{12} = 22$$

Mediana= $$Me = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} = \frac{x_{(\frac{12}{2})} + x_{(\frac{12}{2}+1)}}{2} = \frac{x_6 + x_7}{2} = \frac{15+17}{2} = 16$$

Como se puede observar, en este caso la mediana no es un dato perteneciente a la información, es un parámetro que divide la información dejando el 50% por encima y el 50% por debajo de ella.

. . . . .

5.2.2 La Mediana Cuando la Información se Encuentra Agrupada en Intervalos

Si la información esta agrupada en intervalos iguales, entonces la mediana se calcula según la siguiente expresión:

$$Me = LI + \frac{\frac{n}{2} - fa_{(i-1)}}{f_i} A\ ,$$

Me: Mediana

LI: Límite inferior del intervalo donde se encuentra la mediana (intervalo mediano), el cual se determina observando en que clase se encuentra la posición *n/2*.)

n:	Número de observaciones
$fa_{(i-1)}$:	Frecuencia acumulada anterior al intervalo mediano
f_i:	Frecuencia del intervalo mediano
A:	Amplitud del intervalo

Ejemplo:

Resistencia de 100 Baldosas de la Fabrica "De Las Casas"

Resistencia Kg/cm2	X	f_i	fa	
100 y menos de 200	150	4	4	
200 y menos de 300	250	10	14	
300 y menos de 400	350	21	35	
400 y menos de 500	450	33	68	Interv. Mediano
500 y menos de 600	550	18	86	
600 y menos de 700	650	9	95	
700 y menos de 800	750	5	100	
Sumas		100		

$$\frac{n}{2} = \frac{100}{2} = 50,$$

en la columna de frecuencia acumulada advertimos que la observación número 50 se halla en el cuarto intervalo 4.

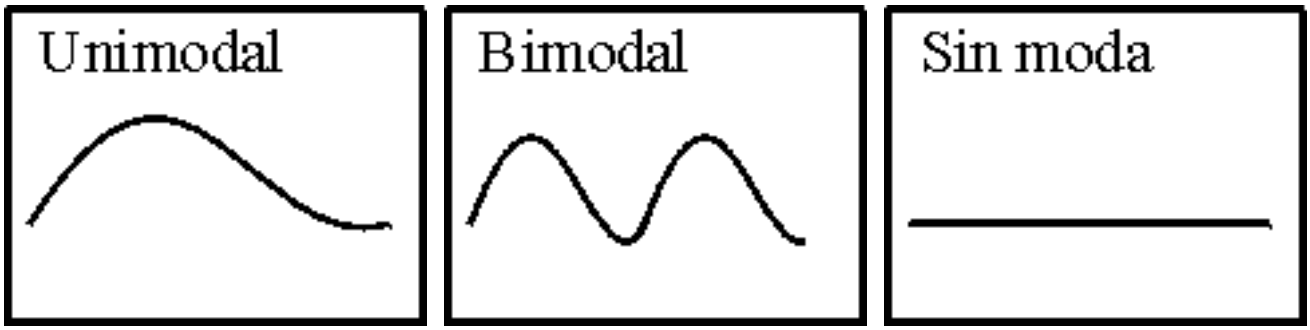
$$Me = LI + \frac{\frac{n}{2} - fa_{(i-1)}}{f_i} A \rightarrow Me = \frac{\frac{100}{2} - 35}{33} 100 = 445.45 \, {Kg}/{Cm^2}$$

Se concluye que el 50% de las baldosas resiste menos de 445.45 Kg/Cm^2 y el 50% resiste mas de 445.45 Kg/Cm^2.

5. Medidas de Tendencia Central

5.3 LA MODA

La moda, como su nombre lo indica, es el valor más común (de mayor frecuencia dentro de una distribución. Una información puede tener una moda y se llama unimodal, dos modas y se llama bimodal, o varias modas y llamarse multimodal. Sin embargo puede ocurrir que la información no posea moda.



5.3.1 La Moda Cuando los datos no están Agrupados en Intervalos

Salario de 50 Operarias de la Fabrica de Confecciones "La Hilacha"

Miles $/día X_i	f_i
50	1
51	3
52	5
53	9
54	12
55	10
56	5
57	3
58	2

El valor que más veces se repite es 54 con una frecuencia de 12, entonces decimos que la moda es *Mo* = 54.000.00 pesos diarios.

Cantidad de Cigarrillos Consumidos por un Fumador en una Semana Dada:

Cantidad X_i	Frecuencia f_i
18	1
19	2
20	1
21	2
22	1

Los valores de mayor frecuencia corresponden a 19 y 21, por lo tanto se trata de una distribución bimodal con Mo_1=19 y Mo_2=21

5.3.2 Cálculo de la Moda Cuando la Información está Agrupada en Intervalos

Cuando la información se encuentra agrupada en intervalos de igual tamaño la moda se calcula con la siguiente expresión.

$$Mo = LI + \frac{f_m - f_{(m-1)}}{2f_m - f_{(m-1)} - f_{(m+1)}} A \,, \quad \text{donde:}$$

. . . . .

Mo: Moda

LI: Límite inferior del intervalo modal

f_m: Frecuencia de la clase modal

$f_{(m-1)}$: Frecuencia de la clase premodal

$f_{(m+1)}$: Frecuencia de la clase posmodal

A : Amplitud de los intervalos

Ejemplo:

Resistencia de 100 Baldosas

Kg/Cm^2	X	fi	
100 y menos de 200	150	4	
200 y menos de 300	250	10	
300 y menos de 400	350	21	Clase premodal
400 y menos de 500	450	33	Clase modal
500 y menos de 600	550	18	Clase posmodal
600 y menos de 700	650	9	
700 y menos de 800	750	5	

$$Mo = LI + \frac{f_m - f_{(m-1)}}{2f_m - f_{(m-1)} - f_{(m+1)}} A$$

$$\rightarrow Mo = 400 + \frac{33-21}{2(33)-21-18} 100 = 444.44 K/Cm^2$$

A pesar que el valor 444.44 no es un dato real de la información asumimos ese parámetro como el de mayor ocurrencia.

CUESTIONARIO Y EJERCICIOS PROPUESTOS

1. ¿Que es una medida de tendencia central?

2. ¿Cuales son las principales medidas de tendencia central?

3. Defina : media aritmética mediana y moda.

4. ¿Cuándo se utiliza la media aritmética ponderada?

5. Enuncie las propiedades de la media aritmética

6. Para cada información de los ejercicios del capitulo 3, calcular e interpretar la media aritmética, la mediana y la moda.

7. La tripulación de un avión, en su itinerario compra los siguientes galones de gasolina:

Ciudad X 200 galones a 4000 pesos el galón

Ciudad Y 250 galones a 3500 pesos el galón
Ciudad Z 300 galones a 3000 pesos el galón

¿Cuál es el costo promedio de la gasolina comprada?

6. Medidas de Posición (Percentiles)

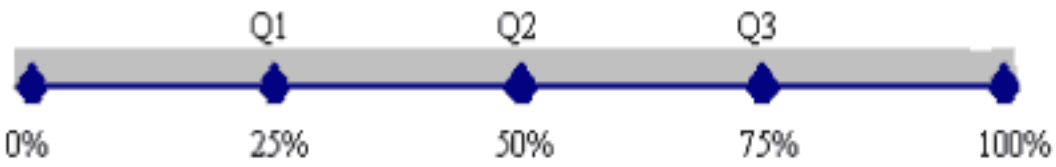
En el Capitulo anterior, vimos lo referente a las medidas de tendencia central, las cuales, a su vez, son también medidas de posición ya que, de todas maneras ocupan un lugar dentro de la información.

Nos ocuparemos ahora de ciertos parámetros posicionales muy útiles en la interpretación porcentual de la información.

6.1 CUARTILES

Las cuartillas o cuartiles son valores posicionales que dividen la información en cuatro partes iguales, el primer cuartil deja el 25% de la información por debajo de él, y el 75% por encima, el segundo cuartil, al igual que la mediana, divide la información en dos partes iguales, y por último el tercer cuartil deja el 75% por debajo de sí, y el 25% por encima.

Gráficamente:



Se necesita, entonces calcular tres cuartillas ya que la cuarta queda automáticamente determinada.

$$Q_k = LI + \frac{\frac{kn}{4} - fa_{(i-1)}}{f_i} A, \quad \text{donde:}$$

k :	Oden del cuartil k = 1,2,3
LI:	Límite inferior del intervalo que contiene el cuartil
$fa_{(i-1)}$:	Frecuencia acumulada hasta el intervalo anterior al que contiene el cuartil
f_i :	Frecuencia del intervalo que contiene el cuartil
n :	Número de observaciones

A : Amplitud de los intervalos

Ejemplo:

Resistencia de 100 Baldosas de la Fabrica "De Las Casas"

Kg/Cm 2	X	f_i	fa	
100 y menos de 200	150	4	4	
200 y menos de 300	250	10	14	
300 y menos de 400	350	21	35	Contiene a Q_1
400 y menos de 500	450	33	68	Contiene a Q_2
500 y menos de 600	550	18	86	Contiene a Q_3
600 y menos de 700	650	9	95	
700 y menos de 800	750	5	100	

Primer cuartil: $k=1 \rightarrow \frac{kn}{4} = \frac{1(100)}{4} = 25$

posición que debe ser ubicada en la frecuencia acumulada, para determinar que clase contiene este cuartil.

$$Q_k = LI + \frac{\frac{kn}{4} - fa_{(i-1)}}{f_i} A \rightarrow Q_1 = 300 + \frac{25-14}{21} 100 = 352.38$$

$$Q_1 = 352.38$$

El 25% de las baldosas resiste menos de 352.38 Kg/Cm^2 y el 75% tiene una resistencia superior.

Como el segundo cuartil es lo mismo que la mediana: $Me=Q_2=445.45Kg/Cm^2$

Calculemos la tercera cuartilla $k=3$

$$Q_k = LI + \frac{\frac{kn}{4} - fa_{(i-1)}}{f_i} A \rightarrow Q_3$$

$$= 500 + \frac{75-68}{18} 100 = 538.88 Kg/cm^2$$

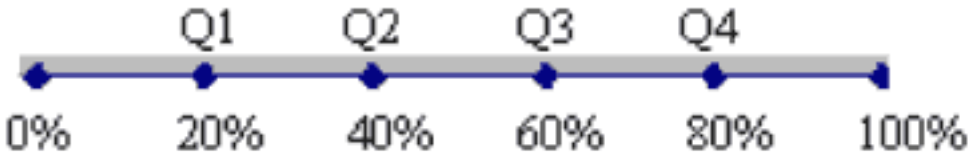
$$Q_3 = 538.88 \text{Kg/Cm}^2$$

El 75% de las baldosas tiene una resistencia inferior a 538..88 Kg/Cm^2 y el 25% una resistencia superior.

6.2 QUINTILES

Los quintiles o quintillas dividen la información en cinco partes iguales, agrupándolas en porcentajes de 20, 40, 60, y 80 por ciento, en consecuencia debemos calcular cuatro parámetros:

Gráficamente:



$$Q_k = LI + \frac{\frac{kn}{5} - fa_{(i-1)}}{f_i} A \quad , \quad k=1,2,3,4.$$

calculemos por ejemplo la segunda quintilla para el ejercicio que traemos:

k=2,

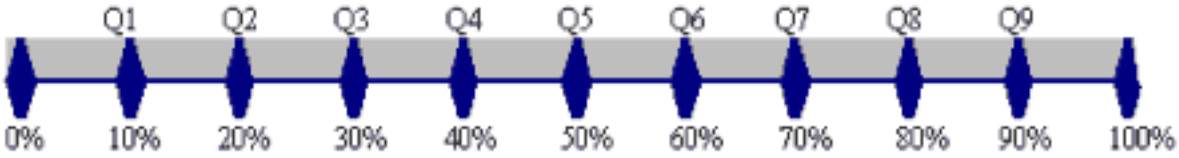
$$\frac{kn}{5} = \frac{2(100)}{5} = 40 \rightarrow Q_2 = \frac{40-35}{33} 100 = 415.15 Kg/cm^2$$

El 40% de las baldosas resiste menos de 415.15kg/cm^2 y el 60% resiste más.

6.3 DECILES

Similarmente, los deciles o decillas dividen la información en diez partes iguales, en cantidades porcentuales de 10 en 10.

$$Q_k = LI + \frac{\frac{kn}{10} - fa_{(i-1)}}{f_i} A \quad , \quad k = 1,2,3,\ldots,9$$



6.4 CENTILES

Obviamente los centiles dividen la información en 100 partes, lo cual facilita la interpretación porcentual de una distribución de frecuencias.

$$Q_k = LI + \frac{\frac{kn}{100} - fa_{(i-1)}}{f_i} A \quad , \quad k=1,2,3,\ldots\ldots,99$$

6.5 RESUMEN

En general para calcular cualquier percentil:

$$Q_k = LI + \frac{\frac{kn}{r} - fa_{(i-1)}}{f_i} A, \quad \text{donde:}$$

r: Número de partes en que se divide la información

k: Orden del percentil $k = 1,2,\ldots\ldots,r-1$

LI:	Límite inferior del intervalo que contiene el percentil
$fa_{(i-1)}$:	Frecuencia acumulada hasta el intervalo anterior al que contiene el percentil
$f_{i:}$	Frecuencia del intervalo que contiene el percentil
n:	Número de observaciones
A:	Amplitud de los intervalos

En nuestro ejercicio, si el gerente de la fabrica de baldosas desea ofrecer un garantía de resistencia mínima. Basado en la muestra que se ha obtenido, si no quiere remplazar ninguna pieza, lógicamente debe afirmar que el producto resiste 100 o más Kg/Cm^2. Pero si esta dispuesto a remplazar el 5% de su producción, entonces:

Se debe dar una garantía de $210kg/cm^2$ de resistencia mínima.

CUESTIONARIO Y EJERCICIOS PROPUESTOS

1. ¿ Para qué se utilizan los percentiles ?

2. ¿ En cuantas partes se divide la información con:

2.1 Los cuartiles
2.2 Los quintiles
2.3 Los deciles
2.4 Los centiles

3. Para la información de los ejercicios 4 y 5 de la sección 3.2 calcular e interpretar ;

3.1 La primera y tercera cuartilla

3.2 El segundo y cuarto quintil
3.3 ¿ Que porcentaje hay entre la primera y tercera quintilla ?
3.4 ¿ Que porcentaje hay entre la primera cuartilla y la segunda quintilla ?
3.5 ¿ Que porcentaje hay entre la tercera cuartilla y el noveno decil ?

7. Medidas de Dispersión

En el análisis estadístico no basta el cálculo e interpretación de las medidas de tendencia central o de posición, ya que, por ejemplo, cuando pretendemos representar toda una información con la media aritmética, no estamos siendo absolutamente fieles a la realidad, pues suelen existir datos extremos inferiores y superiores a la media aritmética, los cuales, en honor a la verdad, no están siendo bien representados por este parámetro.

En dos informaciones con igual media aritmética, no significa este hecho, que las distribuciones sean exactamente iguales, por lo tanto, debemos analizar el grado de homogeneidad entre sus datos. Por ejemplo, los valores 5, 50, 95 tiene igual media aritmética, y mediana que los valores 49, 50,51; sin embargo, para la primera información la media aritmética , se encuentra muy alejada de los valores extremos 5 y 95, cosa que no ocurre con la segunda información que posee igual media aritmética y mediana, vemos entonces que la primera información es mas heterogénea o dispersa que la segunda.

Para medir el grado de dispersión de una variable, se utilizan principalmente los siguientes indicadores:

7.1 Rango o recorrido
7.2 Desviación media
7.3 Varianza y desviación típica o estándar
7.4 Coeficiente de variabilidad.

7.1 RANGO O RECORRIDO

Es la medida de dispersión mas sencilla ya que solo considera los dos valores extremos de una colección de datos, sin embargo, su mayor utilización está en el campo de la estadística no paramétrica.

$$R = Xmax - Xmin$$

Xmax, Xmin son el máximo y el mínimo valor de la variable X, respectivamente.

En el ejemplo introductorio, vemos que el rango para la primera información es R_1=95-5=90, mientras que R_2=51-49=2, se hace pues manifiesta la gran dispersión de la primera información contra la homogeneidad de la segunda.

7.2 DESVIACIÓN MEDIA

La desviación media, mide la distancia absoluta promedio entre cada uno de los datos, y el parámetro que caracteriza la información. Usualmente se considera la desviación media con respecto a la media aritmética:

$$DM = \frac{\sum_{1}^{m} \left| x_i - \overline{X} \right| f_i}{n}$$, donde

DM :	Desviación media
x_i :	Diferentes valores de la variable X
f_i :	Número de veces que se repite la observación x_i
$\overline{X}$	Media aritmética de la información
n :	Tamaño de la muestra.
m :	Número de agrupamientos o intervalos

Ejemplo:

Salario de 50 Operarias de la Fabrica de Confecciones "La Hilacha"

Miles $/día Xi	fi	$\left\|x_i - \overline{X}\right\|$	$\left\|x_i - \overline{X}\right\| f_i$
50	1	4.1	4.1
51	3	3.1	9.3
52	5	2.1	10.5
53	9	1.1	9.9
54	12	0.1	1.2
55	10	0.9	9.0
56	5	1.9	9.5
57	3	2.9	8.7
58	2	3.9	7.8
Sumas	50		70

$$DM = \frac{\sum_{1}^{m} |x_i - \overline{X}| f_i}{n} = \frac{70}{50} = 1.4$$

1.400.00 es el error promedio que se comete al remplazar los ingresos diarios de cada una de las 50 obreras por 54.100 pesos.

7.3 VARIANZA

El problema de los signos en la desviación media, es eludido tomando los valores absolutos de las diferencias de los datos con respecto a la media aritmética. Ahora bien, la varianza obvia los signos elevando las diferencias al cuadrado, lo cual resulta ser más elegante, aparte de que es supremamente útil en el ajuste de modelos estadísticos que generalmente conllevan formas cuadráticas.

La varianza es uno de los parámetros más importantes en estadística paramétrica, se puede decir que, teniendo conocimiento de la varianza de una población, se ha avanzado mucho en el conocimiento de la población misma.

Numéricamente definimos la varianza, como desviación cuadrática media de los datos con respecto a la media aritmética:

$$S^2 = \frac{\sum_{1}^{m} (x_i - \overline{X})^2 f_i}{n}$$, donde:

S^2 : Varianza
x_i : Valor de la variable X
$\overline{X}$: Media aritmética de la información
f_i : Frecuencia absoluta de la observación x_i
n : Tamaño de la muestra.
m : Número de agrupamientos o intervalos

Salario/dia de 50 Operarias en la Fábrica de Confecciones "La Hilacha" (Miles de Pesos)

Miles $ X_i	f_i	$x_i f_i$	$(x_i - \overline{X})$	$(x_i - \overline{X})^2$	$(x_i - \overline{X})^2 f_i$
50	1	50	50-54.1= -4.1	16.81	16.81
51	3	153	51-54.1= -3.1	9.61	28.83
52	5	260	52-54.1= -2.1	4.41	22.05
53	9	477	53-54.1= -1.1	1.21	10.89
54	12	648	54-54.1= -0.1	0.01	0.12
55	10	550	55-54.1= 0.9	0.81	8.10
56	5	280	56-54.1= 1.9	3.61	10.05
57	3	171	57-54.1= 2.9	8.41	25.23
58	2	116	58-54.1= 3.9	15.21	30.42
Sumas	50				160.50

$$S^2 = \frac{\sum_1^m (x_i - \overline{X})^2 f_i}{n} = \frac{160.50}{50} = 3.21$$

Como los datos están expresados en miles de pesos y la varianza se encuentra en forma cuadrática obtenemos una varianza de 3’210.000 pesos. Sin embargo para una mejor comprensión debemos recurrir a la desviación típica o estándar definida como la raíz cuadrada de la varianza:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_1^m (x_i - \overline{X})^2 f_i}{n}} \rightarrow S = \sqrt{3.21} = 1.791$$

El error estándar es de 1.791 pesos/diarios.

En el ejemplo de las baldosas:

Resistencia de 100 Baldosas de La Fábrica “De Las Casas”

Kg/Cm²	X	f_i	$x_i f_i$	$x_i - \bar{x}$	$(x_i - \overline{X})^2 f_i$
100 y menos de 200	150	4	600	-298	355216
200 y menos de 300	250	10	2500	-198	392040
300 y menos de 400	350	21	7350	-98	201684
400 y menos de 500	450	33	14850	2	132
500 y menos de 600	550	18	9900	102	187272
600 y menos de 700	650	9	5850	202	367236
700 y menos de 800	750	5	3750	302	456020
			44800		1959600

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_1^m (x_i - \overline{X})^2 f_i}{n}} \rightarrow S = \sqrt{19596} = 140 kg/cm^2$$

7.4 COEFICIENTE DE VARIABILIDAD

Generalmente interesa establecer comparaciones de la dispersión, entre diferentes muestras que posean distintas magnitudes o unidades de medida.

El coeficiente de variabilidad tiene en cuenta el valor de la media aritmética, para establecer un número relativo, que hace comparable el grado de dispersión entre dos o mas variables, y se define como:

$$CV = \frac{S}{\overline{X}} 100$$

Comparemos la homogeneidad de las dos informaciones anteriores, las cuales tienen diferente unidad de medida.

para el salario:

$$CV_s = \frac{1.791 pesos/dia}{54.1 pesos/dia} = 0.033 \rightarrow CV = 3.3\%$$

para la resistencia

$$CV_R = \frac{140 Kg/cm^2}{448 Kg/cm^2} = 0.3125 \rightarrow CV = 31.25\%$$

Concluimos que es mucho más dispersa la información correspondiente a la resistencia de las baldosas.

CUESTIONARIO Y EJERCICIOS PROPUESTOS

1. ¿ Cuál es la utilidad de las medidas de dispersión?

2. ¿ Cuales son las principales medidas de dispersión?

3. ¿ Cuál es la medida adecuada para comparar la dispersión entre varias variables que posean diferente magnitud o diferente unidad de medida?

4. Para cada una de las informaciones de los ejercicios de los capítulos anteriores, calcular e interpretar:

4.1 Rango
4.2 Desviación media
4.3 Coeficiente de variabilidad

8. Regresión y Correlación Lineal

Hasta ahora hemos hecho la tabulación y el análisis para una sola variable. Pero los investigadores, además de analizar una información en forma individual, generalmente se interesan en establecer cruces y buscar relaciones entre diferentes variables.

8.1 TABLAS DE DOBLE ENTRADA

Para la presentación bidimensional de las variables "X, Y" se procede de la siguiente manera:

- Se ordenan las variables "X, Y" respectivamente
- Se tabulan los valores X horizontalmente, y los valores Y verticalmente.
- Se buscan las frecuencias para cada par ordenado (x_i,y_j).
- Se suma horizontalmente para obtener las frecuencias de "Y" fy_j, y verticalmente para obtener las frecuencias de "X" fx_i .

x_i : Valores de la variable X, $i=1,2,....m$
y_j : Valores de la variable Y, $j=1,2,... k$
fx_i : Frecuencia de la observación x_i
fy_j : Frecuencia de la observación y_j
f_{ij} : Frecuencia conjunta de los valores (x_i,y_j)
$fa0x_i$: Frecuencia acumulada de la variable "X", en el item i
fay_j : Frecuencia acumulada de la variable "Y", en el item j
frx_i : Frecuencia relativa para la variable "X", en el item i
fry_j : Frecuencia relativa para la variable "Y", en el item j
$frax_i$: Frecuencia relativa acumulada para la variable "X"
$fray_j$: Frecuencia relativa acumulada para la variable "Y".

Tabla de Doble Entrada para la Representación de dos Variables "X, Y"

X / Y	X_1	X_2	.....	x_i	.....	x_m	fy_j	fay_j	fry_j	$fray_j$
Y_1	F_{11}	F_{21}	.....	f_{i1}	....	f_{m1}	Fy_1	fay_1	fry_1	$Fray_1$
Y_2	F_{12}	F_{22}	.....	f_{i2}	....	f_{m2}	Fy_2	fay_2	fry_2	$Fray_2$
:	:	:	:	:	:	:	:	:	:	:
y_j	F_{1j}	F_{2j}	....	f_{ij}	....	f_{mj}	fy_j	fay_j	fry_j	$fray_j$
:	:	:	:	:	:	:	:	:	:	:
y_k	f_{1k}	F_{2k}	....	f_{ik}	....	f_{mk}	fy_k	n	fry_k	1.00
fx_i	Fx_1	fx_2	....	fx_i	....	fx_m	n		1.00	
fax_i	fax_1	fax_2	....	fax_i	....	n				
frx_i	Frx_1	frx_2	....	frx_i	....	frx_m	1.00			
$frax_i$	$frax_1$	$frax_2$	....	$frax_i$	....	1.00				

Como se puede advertir en la disposición de las frecuencias, la interpretación de la variable "Y", puede hacerse analizando los relativos propios en forma horizontal, en tanto que el análisis de la variable "X" se hace en forma vertical.

Experiencia Laboral y Salario Diario de 50 Obreras de la Fábrica de Confecciones "La Hilacha". "X" : Experiencia en Años, "Y": Salario Miles de Pesos

X / Y	2	3	4	5	6	7	8	9	fy_j	fay_j	fry_j	$fray_j$
50	1								1	1	0.02	0.02
51		3							3	4	0.06	0.08
52			5						5	9	0.10	0.18
53				8	1				9	18	0.18	0.36
54				2	9	1			12	30	0.24	0.60
55					3	7			10	40	0.20	0.80
56						3	2		5	45	0.10	0.90
57							2	1	3	48	0.06	0.96
58							1	1	2	50	0.04	1.00
fx_i	1	3	5	10	13	11	5	2	50		1.00	
fax_i	1	4	9	19	32	43	48	50				
frx_i	0.02	0.06	0.10	0.20	0.26	0.22	0.10	0.04	1.00			
$frax_i$	0.02	0.08	0.18	0.38	0.64	0.86	0.96	1.00				

Analizando los relativos para cada una de las variables podemos sacar, entre otras, las siguientes conclusiones:

- El 64% tiene una experiencia igual o inferior a 6 años.
- El 68% tiene una experiencia entre 5 y 7 años incluyendo sus extremos.
- El 60% gana 54.000 pesos diarios o menos.
- El 62% gana entre 53.000 y 55.000 pesos incluyendo sus extremos.

Las tablas de doble entrada también pueden usarse para variables cualitativas, o combinarse variables cualitativas con cuantitativas.

Estado Civil y Número de Hijos de
50 Obreras de la Fabrica "La Hilacha"
X: Estado Civil, Y : Número De Hijos.

X / Y	Casada	Soltera	Unión libre	Viuda	f_i	fa	fr	Fra
0		11			11	11	0.22	0.22
1		9	2	1	12	23	0.24	0.46
2	4	9	2	2	17	40	0.34	0.80
3	5	1	1	2	9	49	0.18	0.98
4	1				1	50	0.02	1.00
Total	10	30	5	5	50		1.00	
%	0.20	0.60	0.10	0.10	1.00			

Se deja al lector la interpretación y análisis de esta tabla.

8.2 CORRELACIÓN

En el análisis conjunto para dos o más variables es básica la búsqueda del tipo y grado de la relación que pueda existir entre ellas, o si por el contrario, las variables sean independientes entre sí y la relación que puedan mostrar se debe únicamente al azar, o a través de terceras variables.

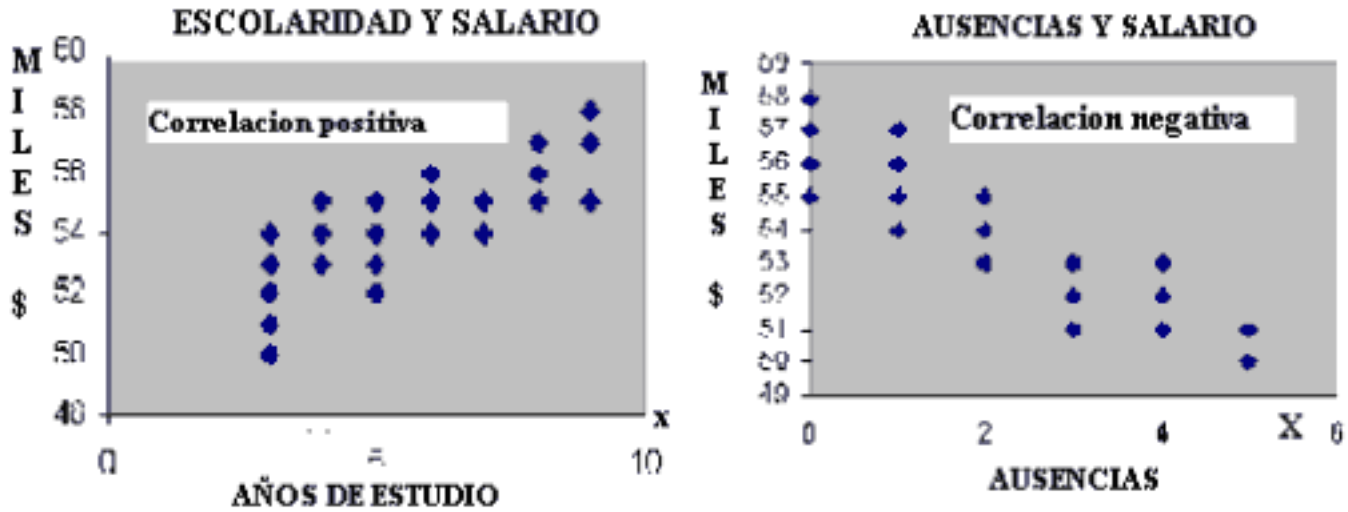
El sondeo del tipo y grado de la correlación, parte desde la misma presunción del investigador, teniendo presente que la búsqueda de relaciones entre variables debe ser lógica, es decir relacionar lo que sea razonable y no datos cuya asociación sea desde cualquier punto de vista absurda.

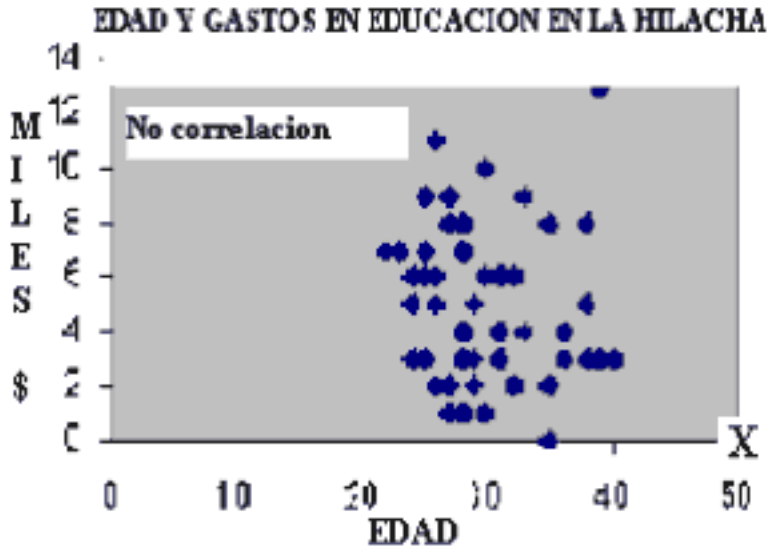
Veamos algunas variables susceptibles de relacionar:

- El peso y estatura de un grupo de adultos.
- Edad y peso de un grupo de niños.
- Ingresos y gastos de arrendamiento de un grupo de familias.
- Escolaridad e ingreso mensual de un grupo de empleados.
- Ventas y utilidades de un almacén de variedades.

En el cuestionario aplicado a las obreras de la "Hilacha", parece que se indaga por ciertas variables que puedan explicar el salario devengado por ellas; como podría ser, los años de experiencia, los años de estudio, las ausencias al trabajo, la evaluación del desempeño por parte de su supervisor, amén de otras variables que pueden tener influencia en la asignación salarial.

Para fortalecer el indicio de correlación inicial, se grafica cada uno de los pares ordenados de las variables (x_i,y_j) en un plano cartesiano, para observar la “nube de puntos” o diagrama de dispersión, donde se advierte la tendencia o no, de la información representada.





A pesar de la ilustración visual que ofrecen las gráficas, solo podemos percibir la tendencia, mas no el grado o fortaleza de la relación, entre la variable independiente “X” y la variable dependiente “Y”.

Para cuantificar la calidad de la dependencia, entre las dos variables, el indicador mas acostumbrado es el ***Coeficiente de correlación,*** definido como:

$r = \frac{S_{x,y}}{S_x S_y}$, donde:

r : Coeficiente de correlación entre “X” y “Y”
S_x: Desviación típica de “X”
S_y: Desciacion típica de “Y”
$S_{x,y}$: Covarianza entre “X” y “Y”

$$S_{x,y} = \frac{\sum_{1}^{m}\sum_{1}^{k}(x_i - \overline{X})(y_j - \overline{Y})f_{ij}}{n} \qquad i=1,2,.....m, \quad ; \quad j=1,2,...k$$

$$r = \frac{S_{x,y}}{S_x S_y} = \frac{\dfrac{\sum_{1}^{m}\sum_{1}^{k}(x_i - \overline{X})(y_j - \overline{Y})f_{ij}}{n}}{\sqrt{\dfrac{\sum_{1}^{m}(x_i - \overline{X})^2 f_i}{n}}\sqrt{\dfrac{\sum_{1}^{k}(y_j - \overline{Y})^2 f_j}{n}}}$$

$$= \frac{1}{n}\frac{\sum_{1}^{m}\sum_{1}^{k}(x_i - \overline{X})(y_j - \overline{Y})f_{ij}}{\sqrt{\dfrac{\sum_{1}^{m}(x_i - \overline{X})^2 f_i \sum_{1}^{k}(y_i - \overline{Y})^2 f_j}{n^2}}}$$

$$r = \frac{\sum_{1}^{m}\sum_{1}^{k}(x_i y_j - \overline{Y}x_i - \overline{X}y_j + \overline{X}\overline{Y})f_{ij}}{\sqrt{\sum_{1}^{m}(x_i^2 - 2\overline{X}x_i + \overline{X}^2)f_i \sum_{1}^{k}(y_j^2 - 2\overline{Y}y_j + \overline{Y}^2)f_j}}$$

$$r = \frac{\sum_{1}^{m}\sum_{1}^{k}x_i y_j f_{ij} - n\overline{X}\overline{Y} - n\overline{X}\overline{Y} + n\overline{X}\overline{Y}}{\sqrt{\left[\sum_{1}^{m}x_i^2 f_i - 2n\overline{X}^2 + n\overline{X}^2\right]\left[\sum_{1}^{k}y_j^2 f_j - 2n\overline{Y}^2 + n\overline{Y}^2\right]}}$$

$$r = \frac{\sum_{1}^{m}\sum_{1}^{k}x_i y_j f_{ij} - n\overline{X}\overline{Y}}{\sqrt{\left[\sum_{1}^{m}x_i^2 f_i - n\overline{X}^2\right]\left[\sum_{1}^{k}y_j^2 f_j - n\overline{Y}^2\right]}}$$

$$r = \frac{\sum_1^m \sum_1^k x_i y_j f_{ij} - n \frac{\sum_1^m x_i f_i}{n} \frac{\sum_1^k y_j f_j}{n}}{\sqrt{\left[\sum_1^m x_i^2 f_i - n\left(\frac{\sum_1^m x_i f_i}{n}\right)^2\right]\left[\sum_1^k y_j^2 f_j - n\left(\frac{\sum_1^k y_j f_j}{n}\right)^2\right]}}$$

$$r = \frac{n\sum_1^m \sum_1^k x_i y_j f_{ij} - \sum_1^m x_i f_i \sum_1^k y_j f_j}{\sqrt{\left[n\sum_1^m x_i^2 f_i - \left(\sum_1^m x_i f_i\right)^2\right]\left[n\sum_1^k y_j^2 f_j - \left(\sum_1^k y_j f_j\right)^2\right]}}$$

En la práctica, cuando no tenemos la información agrupada en una tabla de doble entrada, asumimos que cada observación bivariada tiene frecuencia unitaria, entonces ***r*** se convierte en:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

Tabla de Trabajo para el Calculo de L Coeficiente de Correlacion

X_i	Y_i	X_iY_i	X_i^2	Y_i^2
x_1	y_1	X_1y_1	x_1^2	y_1^2
x_2	y_2	x_2y_2	x_2^2	y_2^2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
x_n	y_n	x_ny_n	x_n^2	y_n^2
$\sum X$	$\sum Y$	$\sum XY$	$\sum X^2$	$\sum Y^2$

El coeficiente de correlación, es un indicador del grado de la relación entre las dos variables, el cual oscila en el intervalo cerrado $[-1,+1]$, es decir, $-1 \le r \le 1$.

Cuando *r* toma un valor extremo, ya sea *r=1 ó r=-1* existe una correlación perfecta positiva o negativa según el signo, como lo podemos corroborar en el siguiente ejemplo:

Aspiración Salarial, de Acuerdo a La Experiencia de las Obreras de la Fabrica de Confecciones "La Hilacha"

Experienc Años	0	1	2	3	4	5	6	7	8	9	10
Miles $/dia	56	58	60	62	64	66	68	70	72	74	76

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

$$= \frac{11(3850) - 55(726)}{\sqrt{\left[11(385) - (55)^2\right]\left[11(48356) - (726)^2\right]}}$$

$$r = \frac{2420}{\sqrt{1210(4840)}} = \frac{2420}{2420} = 1$$, Correlación perfecta positiva

Sin embargo, no todas las relaciones son tan ideales, en el común de los casos $-1 < \boldsymbol{r} < 1$. Empíricamente se afirma que:

1. Si $r = \pm 1$ Correlación perfecta
2. Si $0.9 \le r < 1$ ó $-1 < r \le -0.9$ Correlación excelente
3. Si $0.8 \le r < 0.9$ ó $-0.9 < r \le -0.8$ Correlación buena
4. Si $0.6 \le r < 0.8$ ó $-0.8 < r \le -0.6$ Correlación regular
5. Si $0.3 \le r < 0.6$ ó $-0.6 < r \le -0.3$ Correlación mala
6. Si $-0.3 < r < 0.3$ No hay correlación

Existen desde luego, pruebas estadísticas que miden la bondad de un coeficiente de correlación con un determinado nivel de confiafilidad, pero no son tema de este curso.

Salario Actual y Años de Experiencia de 50 Obreras de la Fabrica "La Hilacha"

Exp	Mil/dia				Exp	Mil/dia			
Años X	Y	XY	X2	Y2	Años X	Y	XY	X2	Y2
4	52	208	16	2704	8	57	456	64	3249
5	54	270	25	2916	6	54	324	36	2916
7	55	385	49	3025	6	55	330	36	3025
6	54	324	36	2916	5	53	265	25	2809
5	53	265	25	2809	7	55	385	49	3025
7	56	392	49	3136	8	56	448	64	3136
5	54	270	25	2916	5	53	265	25	2809
9	58	522	81	3364	9	57	513	81	3249
3	51	153	9	2601	6	54	324	36	2916
6	54	324	36	2916	5	53	265	25	2809
7	54	378	49	2916	2	50	100	4	2500
3	51	153	9	2601	6	55	330	36	3025
6	54	324	36	2916	4	52	208	16	2704
7	55	385	49	3025	5	53	265	25	2809
6	54	324	36	2916	6	54	324	36	2916

8	56	448	64	3136	4	52	208	16	2704
4	52	208	16	2704	8	57	456	64	3249
6	54	324	36	2916	7	56	392	49	3136
5	53	265	25	2809	3	51	153	9	2601
7	55	385	49	3025	8	58	464	64	3364
7	55	385	49	3025	6	55	330	36	3025
7	55	385	49	3025	5	53	265	25	2809
4	52	208	16	2704	6	54	324	36	2916
7	55	385	49	3025	6	53	318	36	2809
5	53	265	25	2809	7	56	392	49	3136
			TOTAL		**294**	**2705**	**16039**	**1850**	**146501**

Se vislumbra una relación positiva, con coeficiente de correlación:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

$$r = \frac{50(16039) - 294(2705)}{\sqrt{\left[50(1850 - (294)^2\right]\left[50(146501) - (2705)^2\right]}} = 0.957$$

Entre la experiencia y el salario actual hay una excelente correlación positiva.

Si escudriñamos en la magnitud de las relaciones entre las diferentes variables cuantitativas, que se han indagado a las obreras de “LA HILACHA” encontramos los siguientes coeficientes de correlación:

	Mil/Dia	Esco	Expe	Califi	Ausen	Edad	Gastos	Hijos
Mil/dia		0.82	0.95	0.86	-0.87	0.59	-0.45	-0.5
Esco	0.82		0.8	0.72	-0.61	0.53	-0.55	-0.72
Expe	0.95	0.8		0.84	-0.84	0.52	-0.43	-0.5
Calif	0.86	0.72	0.84		-0.75	0.56	-0.39	-0.49
Ausen	-0.87	-0.61	-0.84	-0.75		-0.46	0.33	0.34
Edad	0.59	0.53	0.52	0.56	-0.46		-0.08	-0.14
Gastos	-0.45	-0.55	-0.43	-0.39	-0.33	-0.08		0.87
Hijos	-0.5	-0.72	-0.5	-0.49	0.34	-0.14	0.87	

En el problema que nos ocupa, la variable salario/día tiene una excelente correlación positiva, con los años de experiencia, y una buena correlación directa con la calificación y la escolaridad, empero hay una buena relación inversa, con la variable ausencias al trabajo.

8. Regresión y Correlación Lineal

8.3 REGRESIÓN LINEAL

Teniendo ya conocimiento de la intensidad de la correlación entre las variables, manifestada a través del diagrama de dispersión, y el coeficiente de correlación, podemos ensayar el ajuste de un modelo estadístico que se adapte mejor a las n observaciones; lo que lleva por nombre regresión. Uno de los procedimientos muy comunes en el ajuste regresivo es el método de los mínimos cuadrados, que produce estimaciones con menor error cuadrático promedio

8.3.1 Ajuste Rectilíneo (Método de los Mínimos Cuadrados)

La forma general de una ecuación de línea recta es:

$Y = a + bX$ con:

X : Variable independiente
Y : Variable dependiente
a : Término independiente o intercepto
b : Coeficiente de X

Debemos establecer los parámetro "a" y "b" de la ecuación para poder expresar los valores de la variable Y en función de los valores de la variable X, esto es:

$$y_1 = a + bx_1 \ , \ y_2 = a + bx_2 \ ,$$
$$y_3 = a + bx_3 \ , \ldots\ldots\ldots\ , \ y_n = a + bx_n$$

multipliquemos cada una de estas ecuaciones por su respectivo valor de X

$y_1 = a + bx_1$ multiplicado por x_1	$x_1y_1 = ax_1 + bx_1^2$
$y_2 = a + bx_2$ multiplicado por x_2	$x_2y_2 = ax_2 + bx_2^2$
$y_3 = a + bx_3$ multiplicado por x_3	$x_3y_3 = ax_3 + bx_3^2$
. . .	. . . .
. . .	. . . .
. . .	. . . .
$y_n = a + bx_n$ multiplicado por x_n	$x_ny_n = ax_n + bx_n^2$
Sumas $\sum Y = a + b\sum X$ (1)	$\sum XY = a\sum X + b\sum X^2$ (2)

Las ecuaciones (1) y (2) son llamadas ecuaciones normales de la línea recta, de donde se pueden despejar los parámetros *a*, *b* en función de los datos originales.

De (1) tenemos:

$$a = \frac{\sum Y - b\sum X}{n} \quad (3)$$

Remplazando (3) en (2):

$$\sum XY = \left[\frac{\sum Y - b\sum X}{n}\right]\sum X + b\sum X^2$$

$$\sum XY = \frac{\sum X\sum Y}{n} - \frac{b(\sum X)^2}{n} + b\sum X^2$$

$$\sum XY - \frac{\sum X\sum Y}{n} = b\left[(\sum X)^2 - \frac{\sum X^2}{n}\right]$$

$$b = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}$$

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

Las estimaciones para los parámetros son:

$$\hat{b} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} \quad , \quad \hat{a} = \frac{\sum Y - \hat{b}\sum X}{n}$$

El gorrito " ^ " colocado sobre el parámetro indica estimaciones fundamentadas, en los datos muestrales.

Para ajustar el modelo rectilíneo a los ingresos diarios actuales explicados por los años de experiencia, en la "Hilacha", aprovechamos los totales ya calculados en el coeficiente de correlación:

$$\hat{b} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{50(16039) - 294(2705)}{50(1850) - (294)^2} = 1.1$$

$$\hat{a} = \frac{\sum Y - \hat{b}\sum X}{n} = \frac{2705 - 1.1(294)}{50} = 47.63$$

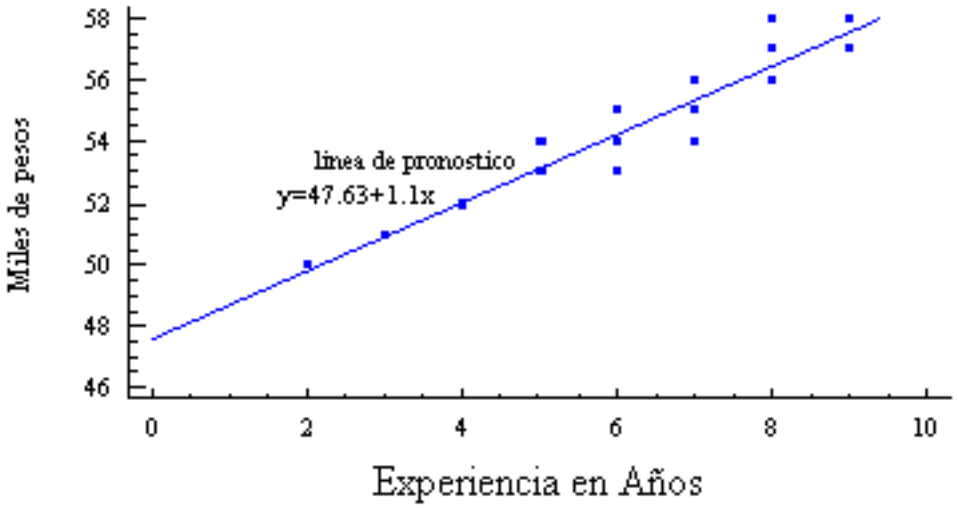
$$\hat{y} = \hat{a} + \hat{b}x \rightarrow \hat{y} = 47.63 + 1.1x$$

Como quiera que los items de la variable salario están en unidades de mil pesos, la ecuación de pronóstico definitiva es:

$$\hat{y} = [47.63 + 1.1x]1000 pesos$$

Experien	$miles/dia	$miles/dia	real – estim	
X años	Y real	Y estim	Error	Frecuen
2	50	49.83	0.17	1
3	51	50.93	0.07	3
4	52	52.03	-0.03	5
5	53	53.13	-0.13	8
5	54	53.13	0.87	2
6	53	54.23	-1.23	1
6	54	54.23	-0.23	9
6	55	54.23	0.77	3
7	54	55.33	-1.33	1
7	55	55.33	-0.33	7
7	56	55.33	0.67	3
8	56	56.43	-0.43	2
8	57	56.43	0.57	2
8	58	56.43	1.57	1
9	57	57.53	-0.53	1
9	58	57.53	0.47	1

Salario Real y Estimado Vs. Experiencia



Insistimos en la existencia de pruebas estadísticas, que miden la bondad de los parámetros estimados y del modelo en sí, a estas alturas de nuestro documento no tenemos las herramientas para aplicarlas, sin embargo en el mercado hay software estadístico, que calcula los parámetros, ajusta los modelos y efectúa las respectivas pruebas, sin exigir al usuario grandes conocimientos de estadística matemática. Se debe tener cuidado, eso sí, en la interpretación adecuada de los resultados.

El siguiente es el reporte parcial producido por el programa de computador Statgraphics plus :

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: miles_dia
Independent variable: Experiencia
------------------------------------------------------------------------------
                             Standard          T
Parameter       Estimate       Error       Statistic        P-Value
------------------------------------------------------------------------------
Intercept        47.6227      0.291056         163.62         0.0000
Slope            1.10158     0.0478493        23.0219         0.0000
------------------------------------------------------------------------------

                           Analysis of Variance
------------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Val
------------------------------------------------------------------------------
Model                      147.172      1      147.172     530.01       0.0C
Residual                   13.3285     48     0.277677
------------------------------------------------------------------------------
Total (Corr.)               160.5     49

Correlation Coefficient = 0.957578
R-squared = 91.6956 percent
```

El programa calcula:

$\hat{a} = 47.6227$ y

$\hat{b} = 1.10158$ y consecuentemente el modelo

$$salario = 47.6227 + 1.10158 * \exp eriencia \ ,$$

el paquete hace también las pruebas *t student* para la hipótesis nula H_0 : $a=0$ vs la hipótesis alternativa H_1 : $a \neq 0$ y H_0 : $b = 0$ vs H_1 : $b \neq 0$, dado que el valor "p" para ambos casos p= 0.0000, con una confiabilidad superior al 99% se rechazan ambas hipótesis de nulidad, a favor de las hipótesis alternativas. En cuanto al valor p = 0.0000 (para la prueba F) en la tabla de análisis de varianza, también se interpreta la validez del modelo con un nivel de confiabilidad superior al 99%.

De otro lado corrobora una correlación positiva excelente r=0.957578 y un coeficiente de determinación R-cuadrado, de 91.6956% que indica el porcentaje de la variable salario explicado por la variable experiencia.

El coficiente de determinación R^2 viene expresado como:

$$R^2 = 1 - \frac{S_E^2}{S_Y^2} \quad ,con \quad E = Y - \hat{Y} \quad , \quad S_E^2 = \text{var}\,ianza\ de\ E,$$

$$S_Y^2 = \text{var}\,ianza\ de\ Y$$

Aprovechemos este pequeño paréntesis, para decir que hoy la tecnología informática ha hecho posible la formulación y solución de complejos modelos multivariados, que constan de cientos de variables, que en años recientes solo se podían teorizar.

En la búsqueda de las variables que explican la variable salario, en la fabrica “La Hilacha” obtenemos el siguiente reporte del programa Statgraphics plus:

Multiple Regression Analysis

Dependent variable: miles_dia

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	49.0432	0.675648	72.587	0.0000
Experiencia	0.553795	0.101613	5.45006	0.0000
Ausencias	-0.275888	0.0826499	-3.33803	0.0018
escolaridad	0.203862	0.0700693	2.90943	0.0058
calificacion	0.206102	0.0996189	2.0689	0.0447
Gastos educ	-0.0927669	0.0450686	-2.05835	0.0458
No hijos	0.296175	0.151952	1.94913	0.0580
Edad	0.0247821	0.0166935	1.48453	0.1451

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	153.762	7	21.9659	136.91	0.0000
Residual	6.73841	42	0.160438		
Total (Corr.)	160.5	49			

R-squared = 95.8016 percent
R-squared (adjusted for d.f.) = 95.1019 percent
Standard Error of Est. = 0.400548
Mean absolute error = 0.299368
Durbin-Watson statistic = 2.18893 (P=0.1821)
Lag 1 residual autocorrelation = -0.123862

```
The StatAdvisor
---------------
   The output shows the results of fitting a multiple linear
regression model to describe the relationship between miles_dia and 7
independent variables.  The equation of the fitted model is

miles_dia = 49.0432 + 0.553795*Experiencia - 0.275888*Ausencias +
0.203862*escolaridad + 0.206102*calificacion - 0.0927669*Gastos educ +
0.296175*No hijos + 0.0247821*Edad

Since the P-value in the ANOVA table is less than 0.01, there is a
statistically significant relationship between the variables at the
99% confidence level.

     The R-Squared statistic indicates that the model as fitted
explains 95.8016% of the variability in miles_dia.  The adjusted
R-squared statistic, which is more suitable for comparing models with
different numbers of independent variables, is 95.1019%.  The standard
error of the estimate shows the standard deviation of the residuals to
be 0.400548.  This value can be used to construct prediction limits
for new observations by selecting the Reports option from the text
menu.  The mean absolute error (MAE) of 0.299368 is the average value
of the residuals.  The Durbin-Watson (DW) statistic tests the
residuals to determine if there is any significant correlation based
on the order in which they occur in your data file.  Since the P-value
is greater than 0.05, there is no indication of serial autocorrelation
in the residuals.

   In determining whether the model can be simplified, notice that the
highest P-value on the independent variables is 0.1451, belonging to
Edad.  Since the P-value is greater or equal to 0.10, that term is not
statistically significant at the 90% or higher confidence level.
Consequently, you should consider removing Edad from the model.
```

El software, analiza los diferentes valores “p” y descalifica la variable edad, al nivel del 90% de confidencialidad, debido a que p=0.1451 hace que el coeficiente de esta variable no sea significante dentro del modelo.

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: miles_dia
-----------------------------------------------------------------------------
                                     Standard          T
Parameter              Estimate        Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                49.4611      0.622757        79.4228          0.0000
Experiencia            0.539854      0.102584        5.26254          0.0000
escolaridad            0.247549     0.0644741        3.83951          0.0004
Ausencias             -0.277339      0.083793       -3.30981          0.0019
calificacion           0.248477     0.0967687        2.56774          0.0138
No hijos                0.36688      0.146301        2.50771          0.0160
Gastos educ           -0.0994745     0.0454649       -2.18794          0.0342
-----------------------------------------------------------------------------

                         Analysis of Variance
-----------------------------------------------------------------------------
Source            Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                    153.408      6       25.568     155.02        0.0000
Residual                 7.09199     43      0.16493
-----------------------------------------------------------------------------
Total (Corr.)              160.5     49

R-squared = 95.5813 percent
R-squared (adjusted for d.f.) = 94.9648 perce
Standard Error of Est. = 0.406116
Mean absolute error = 0.3129
Durbin-Watson statistic = 2.15264 (P=0.2513)
Lag 1 residual autocorrelation = -0.109705
```

```
The StatAdvisor
---------------
   The output shows the results of fitting a multiple linear
regression model to describe the relationship between miles_dia and 6
independent variables.  The equation of the fitted model is

miles_dia = 49.4611 + 0.539854*Experiencia + 0.247549*escolaridad -
0.277339*Ausencias + 0.248477*calificacion + 0.36688*No hijos -
0.0994745*Gastos educ

Since the P-value in the ANOVA table is less than 0.01, there is a
statistically significant relationship between the variables at the
99% confidence level.

     The R-Squared statistic indicates that the model as fitted
explains 95.5813% of the variability in miles_dia.  The adjusted
R-squared statistic, which is more suitable for comparing models with
different numbers of independent variables, is 94.9648%.  The standar
error of the estimate shows the standard deviation of the residuals to
be 0.406116.  This value can be used to construct prediction limits
for new observations by selecting the Reports option from the text
menu.  The mean absolute error (MAE) of 0.3129 is the average value of
the residuals.  The Durbin-Watson (DW) statistic tests the residuals
to determine if there is any significant correlation based on the
order in which they occur in your data file.  Since the P-value is
greater than 0.05, there is no indication of serial autocorrelation in
the residuals.

   In determining whether the model can be simplified, notice that the
highest P-value on the independent variables is 0.0342, belonging to
Gastos educ.  Since the P-value is less than 0.05, that term is
statistically significant at the 95% confidence level.  Consequently,
you probably don't want to remove any variables from the model.
```

Eliminada la variable edad, encontramos un modelo válido con un nivel de confianza superior al 99% cuyos coeficientes son admitidos con una confiabilidad superior al 95%.

$$salario = 49.4611 + 0.5398\,\text{exp}\,eriencia + 0.2475 escolaridad - 0.2773 ausencias + 0.2484 calificación + 0.3668 No.\,hijos - 0.0994 Gastos\,educación$$

R-cuadrado para este modelo es 95.58% , es decir el porcentaje del salario que está siendo explicado por las variables independientes, es ligeramente menor al R-cuadrado anterior (95.8%), sacrificio insignificante cuando se trata de reducir la complejidad del modelo.

Veamos las estimaciones producidas por la ecuación

hijos	Exp	Esco-	Miles $/dia	Gast	ause	Calif	estim salario	Error
2	4	5	52	5	3	1	52.51	0.51
2	5	5	54	6	2	1	53.23	-0.77
3	7	4	55	8	1	4	55.25	0.25
3	6	4	54	9	1	3	54.36	0.36
1	5	3	53	3	2	2	52.91	-0.09
0	7	8	56	1	1	4	55.84	-0.16
1	5	3	54	2	2	3	53.26	-0.74
0	9	9	58	0	0	5	57.79	-0.21
3	3	3	51	10	3	1	51.35	0.35
3	6	3	54	9	2	2	53.59	-0.41
1	7	6	54	3	2	3	54.98	0.98
2	3	3	51	6	5	1	50.82	-0.18
0	6	7	54	1	1	2	54.55	0.55
0	7	7	55	1	1	3	55.34	0.34
0	6	5	54	2	2	3	53.93	-0.07
0	8	8	56	3	1	4	56.18	0.18
1	4	3	52	2	3	2	52.20	0.20
2	6	4	54	5	2	2	53.87	-0.13

2	5	4	53	5	3	2	53.05	0.05
0	7	9	55	4	2	3	55.26	0.26
0	7	8	55	4	1	3	55.29	0.29
1	7	6	55	4	2	3	54.89	-0.11
2	4	3	52	7	3	1	51.82	-0.18
1	7	6	55	3	1	3	55.26	0.26
3	5	3	53	7	2	2	53.25	0.25
0	8	9	57	3	1	5	56.67	-0.33
4	6	5	54	13	2	3	54.30	0.30
3	6	5	55	8	2	3	54.43	-0.57
3	5	4	53	8	2	2	53.40	0.40
3	7	4	55	9	0	3	55.18	0.18
1	8	6	56	4	0	4	56.23	0.23
2	5	4	53	6	2	2	53.23	0.23
0	9	8	57	2	0	4	57.10	0.10
1	6	5	54	3	1	3	54.47	0.47
2	5	3	53	6	2	3	53.23	0.23
2	2	3	50	7	5	1	50.18	0.18
2	6	5	55	6	0	3	54.82	-0.18
2	4	3	52	6	4	1	51.64	-0.36
2	5	4	53	8	3	1	52.50	-0.50

2	6	4	54	8	1	2	53.85	- 0.15
3	4	3	52	11	4	1	51.51	- 0.49
1	8	9	57	3	0	4	57.07	0.07
0	7	8	56	5	0	4	55.72	- 0.28
2	3	3	51	6	4	1	51.10	0.10
1	8	9	58	3	0	4	57.07	- 0.93
2	6	5	55	4	0	2	54.77	- 0.23
1	5	5	53	2	4	1	52.71	- 0.29
2	6	4	54	3	1	1	54.10	0.10
2	6	5	53	7	3	1	53.39	0.39
1	7	6	56	3	0	3	55.54	- 0.46

8.3.2 Ajuste Parabólico (Método Mínimos Cuadrados)

Suele suceder que al dibujar la nube de puntos correspondiente a *n* observaciones bivariante, se observa una tendencia no rectilínea, pero a la cual se le puede ajustar un modelo teórico conocido.

Dentro de la familia de modelos, es de aplicación común el ajuste regresivo polinomial de grado *s* "$s \geq 2$". Similarmente con el procedimiento seguido en el ajuste rectilíneo, vamos a encontrar las ecuaciones normales par una parábola, de forma general $Y = a + bX + cX^2$
es decir $y_1 = a + bx_1 + cx_1^2$, $y_2 = a + bx_2 + cx_2^2$, ..., $y_n = a + bx_n + cx_n^2$. Si cada una de estas ecuaciones la multiplicamos por su respectivo valor de x, y repetimos la acción tenemos:

$$
\begin{array}{lll}
y_1 = a + bx_1 + cx_1^2 & x_1y_1 = ax_1 + bx_1^2 + cx_1^3 & x_1^2y_1 = ax_1^2 + bx_1^3 + cx_1^4 \\
y_2 = a + bx_2 + cx_2^2 & x_2y_2 = ax_2 + bx_2^2 + cx_2^3 & x_2^2y_2 = ax_2^2 + bx_2^3 + cx_2^4 \\
\cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots \\
y_n = a + bx_n + cx_n^2 & x_ny_n = ax_n + bx_n^2 + cx_n^3 & x_n^2y_n = ax_n^2 + bx_n^3 + cx_n^4
\end{array}
$$

sumando se obtienen las siguientes ecuaciones normales

$$\sum Y = na + b\sum X + c\sum X^2 \quad (1)$$

$$\sum XY = a\sum X + b\sum X^2 + c\sum X^3 \quad (2)$$

$$\sum X^2Y = a\sum X^2 + b\sum X^3 + c\sum X^4 \quad (3)$$

De donde se pueden estimar los parámetros de la parábola "$\hat{a},\hat{b},\hat{c}$".

Ejemplo: En un experimento agropecuario, se toma una muestra de 15 unidades de una variedad de árbol frutal, se observa el rendimiento en frutos de acuerdo con la cantidad de fertilizante utilizado:

Gramos	1	1	2	2	3	3	4	5	5	6	7	8	9	9	10
Frutos	10	15	30	25	40	43	50	55	54	53	51	47	41	35	30

$$579 = 15a + 75b + 505c$$
$$3164 = 75a + 505b + 3915c$$
$$21088 = 505a + 3915b + 32617c$$

Resolviendo se obtienen las siguientes estimaciones de los parámetros:

$$\hat{a} = -5.426, \quad \hat{b} = 20.26, \quad \hat{c} = -1.7$$

$$\hat{y} = -5.426 + 20.26x - 1.7x^2$$

Grs	Frut						Frut	
X	Y	XY	X^2	YX^2	X^3	X^4	Est	Error
1	10	10	1	10	1	1	13.13	- 3.13
1	15	15	1	15	1	1	13.13	1.87
2	30	60	4	120	8	16	28.29	1.71
2	25	50	4	100	8	16	28.29	- 3.29
3	40	120	9	360	27	81	40.05	- 0.05
3	43	129	9	387	27	81	40.05	2.95
4	50	200	16	800	64	256	48.40	1.60
5	55	275	25	1375	125	625	53.35	1.65
5	54	270	25	1350	125	625	53.35	0.65
6	53	318	36	1908	216	1296	54.89	- 1.89
7	51	357	49	2499	343	2401	53.03	- 2.03
8	47	376	64	3008	512	4096	47.77	- 0.77
9	41	369	81	3321	729	6561	39.11	1.89
9	35	315	81	2835	729	6561	39.11	- 4.11
10	30	300	100	3000	1000	10000	27.05	2.95
75	**579**	**3164**	**505**	**21088**	**3915**	**32617**		

El programa Statgraphics produce el siguiente reporte:

Polynomial Regression Analysis

Dependent variable: frutos

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-5.42601	2.26557	-2.39499	0.0338
gramos	20.2616	1.02499	19.7677	0.0000
gramos^2	-1.70144	0.093522	-18.193	0.0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	2736.57	2	1368.28	207.75	0.0000
Residual	79.0348	12	6.58624		
Total (Corr.)	2815.6	14			

R-squared = 97.193 percent
R-squared (adjusted for d.f.) = 96.7251 percen
Standard Error of Est. = 2.56637
Mean absolute error = 2.03772
Durbin-Watson statistic = 2.19516 (P=0.1444)
Lag 1 residual autocorrelation = -0.214945

The StatAdvisor

The output shows the results of fitting a second order polynomia model to describe the relationship between frutos and gramos. The equation of the fitted model is

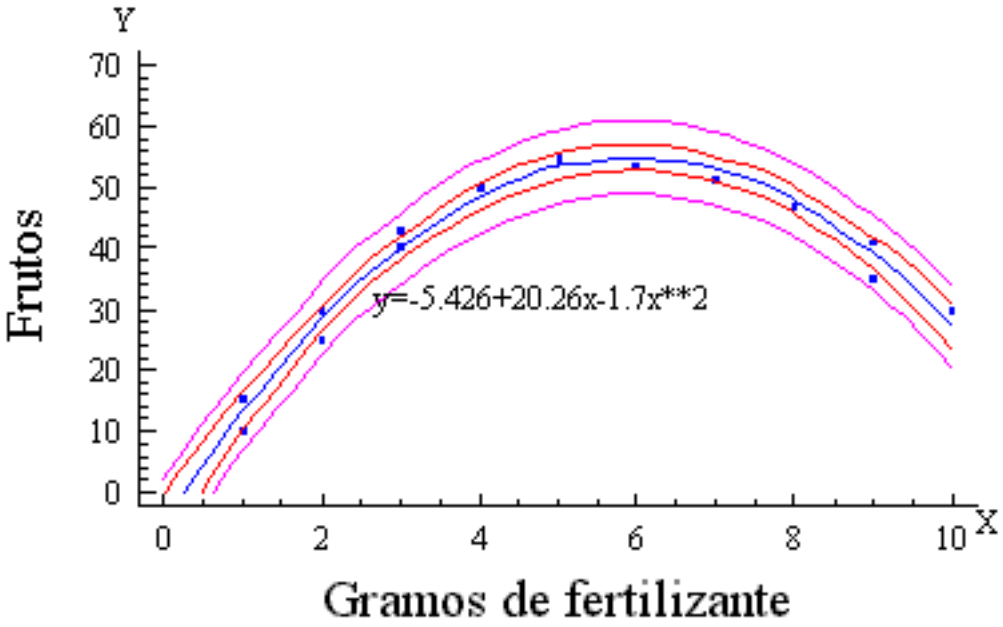
frutos = -5.42601 + 20.2616*gramos-1.70144*gramos^2

Since the P-value in the ANOVA table is less than 0.01, there is a statistically significant relationship between frutos and gramos at the 99% confidence level.

The R-Squared statistic indicates that the model as fitted expla 97.193% of the variability in frutos. The adjusted R-squared statistic, which is more suitable for comparing models with differe numbers of independent variables, is 96.7251%. The standard error the estimate shows the standard deviation of the residuals to be 2.56637. This value can be used to construct prediction limits for new observations by selecting the Forecasts option from the text me The mean absolute error (MAE) of 2.03772 is the average value of th residuals. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the orde in which they occur in your data file. Since the P-value is greate than 0.05, there is no indication of serial autocorrelation in the residuals.

In determining whether the order of the polynomial is appro note first that the P-value on the highest order term of the polynomial equals 4.20114E-10. Since the P-value is less than the highest order term is statistically significant at the 99% confidence level. Consequently, you probably don't want to co any model of lower order.

Parábola Ajustada



CUESTIONARIO Y EJERCICIOS PROPUESTOS

1.

Ingresos y Gastos en Arrendamiento de un Grupo de Familias; en Miles de Pesos.

ing mil	arr mil	ing mil	arr mil	ing mil	Arr Mil	Ing Mil	arr mil	ing mil	arr mil	ing mil	arr mil	ing mil	arr mil
328	118	234	119	236	88	455	154	197	68	240	90	313	125
313	127	415	167	154	76	59	13	366	130	293	98	145	50
455	171	218	73	272	109	301	126	271	103	216	80	206	78
172	84	310	108	324	136	127	79	368	153	454	176	368	140
352	135	231	95	451	155	303	134	344	137	289	115	231	115
265	102	137	63	296	118	106	42	134	49	423	156	371	152
435	142	59	24	166	73	184	75	86	54	391	125	521	197
340	109	400	140	458	179	227	90	277	110	352	130	352	141
367	130	287	96	179	72	531	174	169	68	318	128	269	110
328	119	196	62	217	66	240	90	266	97	347	125	449	133
375	145	325	117	350	161	252	124	589	205	339	111	307	119
251	76	329	133	310	115	240	107	338	106	110	56	260	95
303	110	96	24	334	98	337	106	336	130	184	64	549	222
356	127	420	157	336	93	513	182	345	116	164	78	281	90
148	64	212	81	429	147	325	113	425	124	355	127	225	71
373	134	390	147	307	125	322	126	269	98	416	145	297	105
295	109	456	178	215	100	199	97	353	141	297	127	216	83
321	108	174	64	443	148	74	34	268	86	253	82	325	112
140	59	310	111	203	83	321	106	350	118	148	65	399	144
440	154	376	124	199	97	162	65	399	155	335	111	269	112
414	158	293	133	340	123	384	145	264	129	152	56	347	131
402	137	309	134	270	82	375	143	442	151	295	119	473	155
172	86	294	113	286	120	401	167	239	97	356	123	436	162
301	122	307	140	254	105	321	137	298	105	207	84	358	139
257	93	249	94	540	199	268	111	513	182	361	151	358	129
264	86	422	143	293	104	325	111	514	188	275	93	390	139
406	166	301	117	309	92	290	94	206	66	331	130	448	175
396	158	645	239	329	105	309	116	287	112	349	138	234	89
346	135	363	122	309	128	283	104						

1.1 Calcular el coeficiente de correlación e interpretarlo.
1.2 Ajustar el modelo adecuado para esta información.
1.3 ¿ Cuánto se estima, debe pagar una familia con ingreso mensual de 270.000 pesos?

2. ¿ Que es un coeficiente de correlación ?
3. Cuando hay correlación:

3.1 Perfecta
3.2 Excelente
3.3 Buena

4. ¿ Cuales son las ecuaciones normales de la línea recta ?

9. Tasas e Índices

Como ya se dijo, el análisis de un fenómeno basado en las cifras absolutas, ofrece una idea general de su tendencia o comportamiento; pero para efectos de establecer comparaciones adecuadas del mismo fenómeno con otra región, o su ocurrencia a través del tiempo, se utilizan ciertos indicadores denominados tasas e índices

9.1 TASA

Una tasa es la resultante de una fracción, en donde el numerador está contenido dentro del denominador:

Ejemplos:

$$D = \frac{R}{M}1000$$

D: Tasa de deserción escolar.
R: Número de retiros durante el año.
M: Número total de matriculados durante el año.

$$TE = \frac{PEAO}{PEA}1000$$

TE: Tasa de empleo.
PEAO: Población económicamente activa ocupada.
PEA **:** Población económicamente activa.

Valga anotar que a las tasas se les debe multiplicar por una constante k, la cual generalmente es 100, 1000 o múltiplos de ellos, con el fin de convertirlos en porcentajes, por millares etc.

En demografía, las tasas son de uso frecuente, entre otras, mencionaremos las siguientes:

$$TM = \frac{D}{P}1000$$

Donde:

TM : Tasa de mortalidad.
D : Número de defunciones en un periodo y área dada.
P : Población total en esa área a mitad del periodo.

$$TN = \frac{N}{P}1000$$

Donde

TN : Tasa de natalidad
N : Número de nacidos vivos ocurridos en un periodo y área dada
P : Población total del área a mitad del periodo.

$$TC = \frac{M}{P}1000$$

Donde:

TC : Tasa de nupcialidad.
M : Número de matrimonios efectuados en un periodo y área dada.
P : Total de la población a mitad del periodo.

El siguiente cuadro muestra la evolución de la tasa de desempleo en Colombia, resultados obtenidos de la encuesta nacional de hogares para los periodos comprendidos entre los años 1.990 –2.000

Tasas de Desempleo en Colombia
1.990-2.000

Año	Mes	Tasa	Año	Mes	Tasa
1990	**Marzo**	10.1	**1995**	**septiembre**	8.7
	Junio	10.9		**diciembre**	9.5
	Septiembre	10.2	**1996**	**marzo**	10.2
	Diciembre	10.6		**junio**	11.6
1991	**Marzo**	10.8		**septiembre**	12
	Junio	10.8		**diciembre**	11.3
	Septiembre	9.8	**1997**	**marzo**	12.3
	Diciembre	9.5		**junio**	13.4
1992	**Marzo**	10.8		**septiembre**	12.1
	Junio	11.2		**diciembre**	12
	Septiembre	9.2	**1998**	**marzo**	14.4
	Diciembre	9.8		**junio**	15.9
1993	**Marzo**	9.7		**septiembre**	15
	Junio	9.1		**diciembre**	15.6
	Septiembre	7.8	**1999**	**marzo**	19.5
	Diciembre	7.8		**junio**	19.9
1994	**Marzo**	10.2		**septiembre**	20.1
	Junio	9.9		**diciembre**	18
	Septiembre	7.6	**2000**	**marzo**	20.3
	Diciembre	8		**junio**	20.4
1995	**Marzo**	8.1		**septiembre**	20.5
	Junio	9		**diciembre**[p]	19.7

9.2 ÍNDICE

Un número índice, como comúnmente se le llama, es un indicador de los cambios relativos de una o más variables a través del tiempo.

Entre las principales aplicaciones de los números índice, está la de establecer comparaciones entre los indicadores de las diferentes zonas geográficas, profesiones , grupos étnicos etc.

Para la construcción de un número índice, se procede ante todo, a fijar el periodo de referencia o "periodo base" de la serie temporal, teniendo presente que debe ser un periodo normal, esto es, que no se hayan presentado situaciones fortuitas (guerras, terremotos, incendios u otro tipo de imprevisto), que incidan en el valor de la variable para ese periodo. Además debe considerarse un periodo reciente que haga comparables los diferentes valores de las variables consideradas. 7/30

9.2.1 Índice Simple

Un número índice simple, es aquel que se calcula para una sola variable, dividiendo cada uno de los valores de la serie cronológica, por el valor correspondiente al "periodo base" previamente definido.

9.2.1.1 Índice de Base Fija

$I_p = \frac{p_n}{p_0} 100$, si la variable se refiere a precios

$I_q = \frac{q_n}{q_0} 100$, si la variable se refiere a cantidades

I_p **:** Índice de precios

P_n: Precio del artículo en el periodo n

P_0 : Precio del artículo en el periodo base

I_q *:* Índice de cantidades

q_n **:** Cantidad del articulo en el periodo n

q_0 **:** Cantidad del articulo en el periodo bas

Precio Promedio del Kilovatio/Hora 1995-2001
Pagado por la Fabrica de Confecciones "La Hilacha"

Año	Precio Kw/hora	Índice 1995=100 %	Índice 1998=100 %
1,995	9	1	0.47
1,996	12	1.33	0.63
1,997	15	1.67	0.79
1,998	19	2.11	1
1,999	24	2.67	1.26
2,000	30	3.33	1.58
2,001	37	4.11	1.95

Consumo Promedio de Energía en
La Fabrica de Confecciones "La Hilacha"

Año	cantidad Kw/mes	Indice 1995=100%	Indice 1998=100%
1,995	3,333.33	1.00	0.49
1,996	6,666.67	2.00	0.97
1,997	4,666.67	1.40	0.68
1,998	6,842.11	2.05	1.00
1,999	5,833.33	1.75	0.85
2,000	4,666.67	1.40	0.68
2,001	4,324.32	1.30	0.63

En la primera tabla hemos calculado los índices de precios simples, con base en 1995 y 1998 respectivamente, pero no se han tenido en cuenta las cantidades, mientras que en la segunda tabla se han calculado los índices de cantidades sin considerar los precios.

Calculemos, ahora los índices del valor relativo, que considere tanto los precios como las cantidades:

$$valor\ relativo = \frac{p_n q_n}{p_0 q_0} 100$$

Precio y Consumo Promedio de Energía en La Fabrica de Confecciones "La Hilacha"

Año	Precio Kw/h	Cantidad Kw/mes	pq	Índice 1995=100%	Índice 1998=100%
1,995	9	3,333.33	30,000	1.00	0.23
1,996	12	6,666.67	80,000	2.67	0.62
1,997	15	4,666.67	70,000	2.33	0.54
1,998	19	6,842.11	130,000	4.33	1.00
1,999	24	5,833.33	140,000	4.67	1.08
2,000	30	4,666.67	140,000	4.67	1.08
2,001	37	4,324.32	160,000	5.33	1.23

9.2.1.2 Índice de Base Móvil

Solo hemos considerado, los índices simples de base fija, esto es, con un periodo base determinado. Es común que interese comparar un índice con el índice del periodo inmediatamente anterior, en consecuencia se debe fijar el periodo base en el periodo anterior al referenciado, y así sucesivamente hasta completar la serie, al cual se le nombra índice de base móvil.

Variaciones del Salario Promedio Diario en La Fabrica de Confecciones "La Hilacha"

Año	Miles $/dia	Indice 1995=100%	Índice 100%=año ant.	Variación
1,995	18.70	1.00	-	-
1,996	23.80	1.27	1.27	0.273
1,997	30.80	1.65	1.29	0.294
1,998	38.50	2.06	1.25	0.250
1,999	47.00	2.51	1.22	0.221
2,000	50.20	2.68	1.07	0.068
2,001	54.10	2.89	1.08	0.078

9.2.2 Índices Compuestos (Globales)

Un número índice compuesto, muestra los cambios de un conjunto de variables, auque sus unidades de medidas, cantidades y precios, en el tiempo, sean diferentes entre sí. Cuando hablamos por ejemplo de los índices indicadores del costo de la canasta familiar, se toman en cuenta muchos artículos cuyos consumos inciden en el costo de vida, con una ponderación o importancia diferente en cada caso. Colectivamente no es lo mismo un cambio en el precio de la carne, huevos o leche, que un cambio en el precio de los perfumes, joyas o cualquier otro artículo suntuoso.

9.2.2.1 *Índice de Laspeyres*

Este índice asume como ponderaciones, en el cálculo del índice global, las cantidades de los artículos en el periodo base.

$$P_L = \frac{\sum p_n q_0}{\sum p_0 q_0} 100$$

Donde:

P_L : Índice de precios global (Laspeyres).
q_0 : Cantidad del periodo base.
p_0 : Precio del artículo en el periodo base
p_n : Precio del artículo en el periodo n

Índice de Precios de Cuatro Artículos

						valor de q0 a los precios de			
	Cant	Precios							
	q0	p0	P1	p2	p3	p0	p1	p2	p3
Art Año	1998	1998	1999	2000	2001	1998	1999	2000	2001
A	5	10	12	14	15	50	60	70	75
B	10	20	24	25	25	200	240	250	250
C	15	10	10	11	12	150	150	165	180
D	20	25	27	28	30	500	540	560	600
Sumas						900	990	1045	1105
Indice						**1.00**	**1.10**	**1.16**	**1.23**

9.2.2.2 Índice de Paasche

El estadístico Paasche, sugiere que las ponderaciones sean las cantidades utilizadas en el periodo n. Se obtiene entonces el siguiente indicador:

$$P_P = \frac{\sum p_n q_n}{\sum p_0 q_n} 100$$

Este índice, es poco utilizado debido al dinamismo de q_n , necesitando nuevas ponderaciones cada vez que se cambia de periodo.

9.2.2.3 Índice ideal de Fisher

Se propone el promedio geométrico entre los dos índices anteriores:

$$P_F = \sqrt{P_L P_P} = \sqrt{\left[\frac{\sum p_n q_0}{\sum p_0 q_0}\right]\left[\frac{\sum p_n q_n}{p_0 q_n}\right]} 100$$

Una de las principales aplicaciones de los índices de precios, es la de medir la deflación e inflación, que es la variación que existe en el poder adquisitivo

del dinero. También podemos utilizar, los índices de precios al consumidor para determinar el salario real de un grupo de personas.

$$Salario\ real = \frac{Salario\ no\min al}{índice\ de\ precios\ al\ consumidor} 100$$

Salario Promedio Nominal y Real en la Fabrica "La Hilacha"

Años	Indice 1998=100%	variación Anual	miles/día nominal	miles/día Real
1994	50.10	22.59	18.7	37.32
1995	59.86	19.46	23.8	39.76
1996	72.81	21.63	30.8	42.30
1997	85.69	17.68	38.5	44.93
1998	100.00	16.7	47	47.00
1999	109.27	9.23	50.2	45.94
2000	118.79	7.81	54.1	45.54

Dado el deterioro del salario real en los dos últimos años debería considerarse un generoso aumento.

CUESTIONARIO Y EJERCICIOS PROPUESTOS

1. ¿Qué es una tasa?

2. ¿Qué es un índice?

3. ¿Para qué se utilizan los números índices?

4. ¿Cómo se construye un número índice simple?

5. ¿Cómo se construye un número índice compuesto?

6. Los precios y las cantidades de un articulo X vienen dados en la siguiente tabla:

Año	Precio	Cantidad
1995	1200	400
1996	1500	400
1997	1800	410
1998	2200	430
1999	2600	430
2000	3000	440

Tomando como año base 1995, calcular para los otros años:

6.1 Los índices de precios.
6.2 Los índices de cantidades.
6.3 Los índices de valores.

7. A continuación se relacionan los precios y las cantidades del año base, de cuatro artículos diferentes:

	Cant	Precios				
	q0	p0	p1	p2	p3	p4
Art Año	1997	1997	1998	1999	2000	2001
A	180	200	250	300	350	400
B	100	50	60	70	80	90
C	400	100	120	130	150	180
D	120	20	30	30	40	40

Calcular el índice de Laspeyres

8.

Salario Mínimo Legal Diario en Colombia e Índice de Precios al Consumidor para el Año 2.000-2001 (Periodo Base Diciembre de 1998).

Año	MES	Valor Indice	Salario nominal
2000	1	110.64	260000
2000	2	113.19	260000
2000	3	115.12	260000
2000	4	116.27	260000
2000	5	116.88	260000
2000	6	116.85	260000
2000	7	116.81	260000
2000	8	117.18	260000
2000	9	117.68	260000
2000	10	117.86	260000
2000	11	118.24	260000
2000	12	118.79	260000
2001	1	120.04	286000
2001	2	122.31	286000
2001	3	124.12	286000
2001	4	125.54	286000
2001	5	126.07	286000
2001	6	126.12	286000

Calcular el salario real para cada uno de los meses.

10. Nociones de Probabilidad (Eventos)

"Los planes corresponden al hombre,
las probabilidades a Dios."
Proverbio chino

Todo experimento debe ser susceptible de repeticiones conservando las mismas condiciones con las cuales se realizó su antecesor. Esto es, el investigador debe fijar esas condiciones, bajo las cuales se realizarán las sucesivas repeticiones del experimento y conservarlas en cada una de las réplicas, de tal manera que sus inferencias resulten lo más fiables posible. Sin embargo, aun así no siempre se obtienen los mismos resultados, pues a veces participan factores incontrolables que aparentemente no obedecen a ninguna causa natural, ni intervención humana intencionada y que denominamos ***Azar*** o casualidad.

Desde el punto de vista de la presencia o no de la *contingencia* en los resultados, si definimos experimentos determinísticos y experimentos aleatorios:

Experimento determinístico es aquel en el cual, bajo las mismas condiciones experimentales, las repeticiones del experimento absolutamente todas, siempre producen el mismo resultado.

El experimento Aleatorio, conservando las mismas condiciones experimentales, los resultados no se pueden predecir, con exactitud, para ninguna repetición.

Sí, por ejemplo lanzamos una moneda al aire para observar de cual lado cae, no podemos pronosticar con certeza, si se presenta sello o se presenta cara. Tenemos entonces presente el componente del azar y por consiguiente un experimento aleatorio. No ocurriría igual si la moneda estuviese diseñada igual por ambos lados y por consiguiente sería un experimento determinístico:

Todos los posibles resultados de un experimento aleatorio, conforman el espacio muestral que representaremos por "S", a cualquier subconjunto del espacio muestral se le denomina suceso o evento aleatorio y lo denotaremos con "E". $E \subset S$. Cada uno de los elementos del espacio muestral se denomina evento elemental "e": $e \in E \subset S$

Definiciones sobre Sucesos:

- El evento $A \cup B$ ocurre cuando se verifica uno de los dos, o ambos sucesos.

- El evento $A \cap B$ se presenta cuando ocurren los dos simultáneamente.

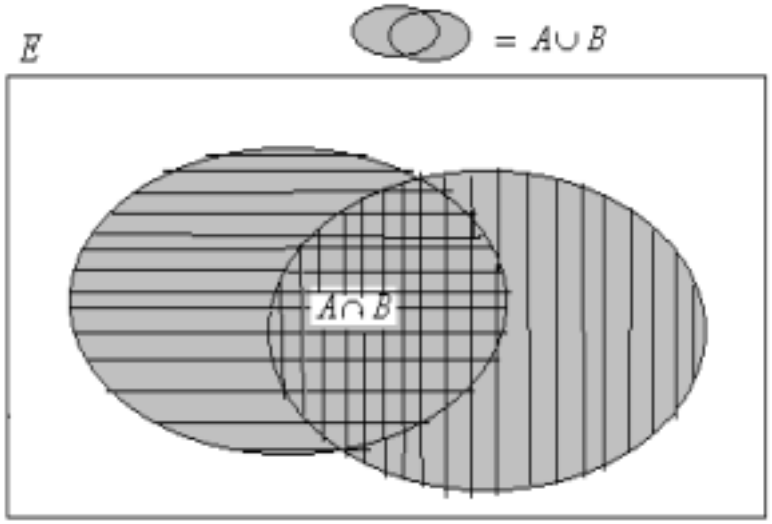
Evento o suceso elemental $E = (e)$

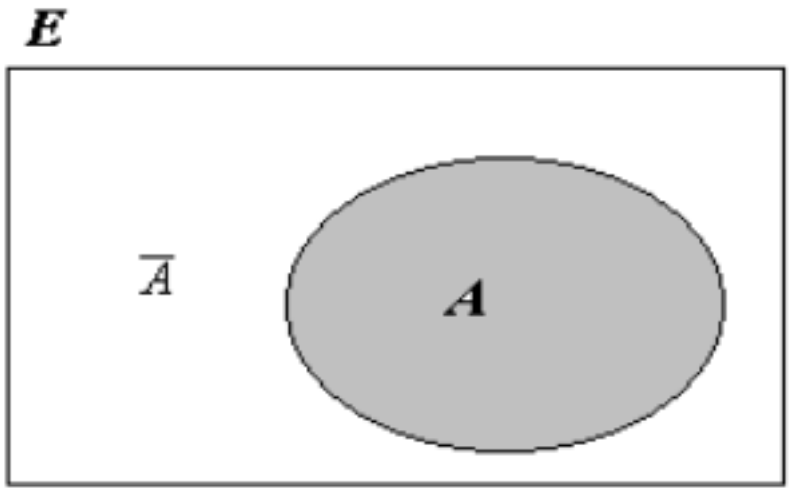
Evento o suceso seguro Siempre se presenta en un experimento: *S*

Evento o suceso imposible nunca ocurre dentro un experimento: •

Eventos incompatibles Dos o más sucesos son incompatibles o excluyentes cuando la ocurrencia de uno impide la presencia de los otros. Si E_1, E_2 excluyentes entonces $E_1 \cap E_2 = \Phi$

Sucesos complementarios o contrarios Dos sucesos son complementarios cuando son mutuamente excluyentes y su unión conforma: el espacio muestral***:*** E *y* $\overline{E}$ son complementarios $\rightarrow E \cap \overline{E} = \Phi$. $E \cup \overline{E} = S$. Si E es un evento seguro, entonces E=S





En general, los sucesos o eventos, tienen las mismas propiedades de los conjuntos.

Propiedades de los eventos:

	Unión	intersección
Conmutativa	$A \cup B = B \cup A$	$A \cap B = B \cap A$
Asociativa	$A \cup (B \cup C) =$ $(A \cup B) \cup C$	$A \cap (B \cap C) =$ $(A \cap B) \cap C$
Ídem potente	$A \cup A = A$	$A \cap A = A$
Simplificación	$A \cup (B \cap A) = A$	$A \cap (B \cup A) = A$
Distributiva	$A \cup (B \cap C) =$ $(A \cup B) \cap (A \cup C)$	$A \cap (B \cup C) =$ $(A \cap B) \cup (A \cap C)$
Elemento neutro	$A \cup \Phi = A$	$A \subset S \rightarrow A \cap S = A$
Absorción	$A \cup S = S$	$A \cap \Phi = \Phi$

- El complemento de la unión de dos sucesos es la intersección de sus complementos: $\overline{(A \cup B)} = \overline{A} \cap \overline{B}$
- El complemento de la intersección de dos sucesos es la unión de sus complementos: $\overline{(A \cap B)} = \overline{A} \cup \overline{B}$

. . . . .

Ejemplo:

Lanzamos una moneda para observar, si cae del lado de cara o del lado de sello:

- Espacio muestral $S = \{cara, sello\ \} = \{c, s\ \}$
- Eventos elementales $E_1 = \{c\ \}$, $E_2 = \{\ s\}$
- Evento seguro $E = S = \{c\ o\ s\}$
- Evento imposible $\{ni\ "cara", ni\ "sello"\} = \Phi$
- E_1 y E_2 son eventos excluyentes.

Ejemplo:

Lanzar un par de dados, marcados c/u con los números 1,2,3,4,5 y 6

Espacio muestral

$$S = \begin{Bmatrix} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{Bmatrix}$$

E_1: (suma igual a 2): $E_1 = \{(1,1)\}$ *suceso elemental*

E_2: (suma igual a 3): $E_2 = \{(1,2),(2,1)\}$

E_3: (suma igual a 4): $E_3 = \{(1,3),(2,2),(3,1)\}$

E_4: (suma igual a 5): $E_4 = \{(1,4),(2,3),(3,2),(4,1)\}$

E_5: (suma igual a 6): $E_5 = \{(1,5),(2,4),(3,3),(4,2),(5,1)\}$

E_6: (suma igual a 7): $E_6 = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$

E_7: (suma igual a 8): $E_7 = \{(2,6),(3,5),(4,4),(5,3),(6,2)\}$

E_8: (suma igual a 9): $E_8 = \{(3,6),(4,5),(5,4),(6,3)\}$

E_9: (suma igual a 10): $E_9 = \{(4,6),(5,5),(6,4)\}$

E_{10}: (suma igual a 11): $E_{10} = \{(5,6),(6,5)\}$

E_{11}: (suma igual a 12): $E_{11} = \{(6,6)\}$ *suceso elemental*

Con la unión e intersección de dos o mas eventos, se generan nuevos sucesos.

Ejemplo:

En una mesa hay un juego (28 fichas) de dominó, se voltea una ficha para observar sus números:

Espacio muestral

$$S = \left\{ \begin{array}{ccccccc} (0,0) & (0,1) & (0,2) & (0,3) & (0,4) & (0,5) & (0,6) \\ & (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ & & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ & & & (3,3) & (3,4) & (3,5) & (3,6) \\ & & & & (4,4) & (4,5) & (4,6) \\ & & & & & (5,5) & (5,6) \\ & & & & & & (6,6) \end{array} \right\}$$

E_1: La diferencia absoluta entre sus componentes sea igual a 0

$$E_1 = \{(0,0) \quad (1,1) \quad (2,2) \quad (3,3) \quad (4,4) \quad (5,5) \quad (6,6)\}$$

E_2: La diferencia absoluta entre sus componentes sea igual a 1

$$E_2 = \{(0,1) \quad (1,2) \quad (2,3) \quad (3,4) \quad (4,5) \quad (5,6)\}$$

E_3: La diferencia absoluta entre sus componentes sea igual a 2

$$E_3 = \{(0,2) \quad (1,3) \quad (2,4) \quad (3,5) \quad (4,6)\}$$

E_4: La diferencia absoluta entre sus componentes sea igual a 3

$$E_4 = \{(0,3) \quad (1,4) \quad (2,5) \quad (3,6)\}$$

E_5: La diferencia absoluta entre sus componentes sea igual a 4

$$E_5 = \{(0,4) \quad (1,5) \quad (2,6)\}$$

E_6: La diferencia absoluta entre sus componentes sea igual a 5

$$E_6 = \{(0,5) \quad (1,6)\}$$

E_7: La diferencia absoluta entre sus componentes sea igual a 6

$$E_7 = \{(0,6)\}$$

10. Nociones de Probabilidad (Eventos)

10.1 NOCIONES DE CONTEO

10.1.1 Principio Fundamental 1

Si un suceso A puede ocurrir de n maneras y otro suceso B puede ocurrir m maneras, entonces el suceso A ó B (Sucede el evento A ó sucede el evento B) puede ocurrir de $n+m$ formas, siempre y cuando los eventos no puedan suceder simultáneamente.

Ejemplo:
En el lanzamiento de un dado, de cuantas maneras se puede obtener un número inferior a 2 o mayor que 4?

A: (número inferior a 2) sucede solo de una manera.
B : (número superior a 4), sucede de dos maneras
A ó B (número inferior a 2 o superior a 4)

sucede de 1+2=3 maneras.

10.1.2 Principio Fundamental 2

Si un seceso A puede suceder de n maneras y un suceso B de m formas, entonces el suceso A y B (sucede el evento A y sucede el Evento B) puede ocurrir de *n(m)* modos.

Ejemplo:
De cuantas maneras distintas pueden caer 2 dados, lanzados simultáneamente:

A: (dado 1) puede caer de 6 maneras.

B : (dado 2) puede caer de 6 maneras
A y B (dado 1 y dado 2) sucede de 6(6) =36 maneras

10.1.3 Permutaciones

Se le llama permutacion a cada uno de los arreglos de n elementos, cuya diferenciación mutua se debe al orden en que están colocados sus elementos. Al total de permutaciones obtenidas con n elementos se le representa por:

$$P_n = n! = n(n-1)(n-2)\ldots(2)(1) \qquad 1$$

Ejemplo:
Cuantas palabras diferentes se pueden formar con las letras n, l, o, e; así no tengan sentido?

$$P_4 = 4! = 4(3)(2) = 24$$

nloe, nleo, nelo, neol, nole noel, lnoe, lneo, leno, leon, lone, loen, elon, elno, enlo, enol, eoln, eonl, olne, olen, oeln, oenl, onle, onel.

10.1.4 Variaciones

A cada uno de los arreglos de *r* elementos obtenidos de un grupo de *n* elementos $(r \leq n)$, cuya diferenciación mutua se deba a los elementos ó el orden de colocación, se le denomina variación. El número total de variaciones se representa por:

$$V_r^n = \frac{n!}{(n-r)!} = \frac{n(n-1)(n-2)\ldots 2(1)}{(n-r)(n-r-1)(n-r-2)\ldots 2(1)}$$

Ejemplo:
Cuantos números de tres cifras se pueden construir con los dígitos 1,2,3,4,5,6,7,8,9,0 si ninguno se puede repetir

$$V_3^{10} = \frac{10!}{(10-3)!} = \frac{10(9)(8)7!}{7!} = 720 \quad \textit{números} \quad \textit{diferentes}$$

10.1.5 Combinaciones

A cada uno de los arreglos de *r* elementos obtenidos de un grupo de *n* elementos $(r \leq n)$, cuya diferenciación mutua se deba a los elementos sin importar el orden de colocación de ellos, se le denomina combinación. El número total de combinaciones se representa por:

$$C_r^n = \frac{n!}{(n-r)!r!} = \frac{n(n-1)(n-2).....3(2)(1)}{(n-r)(n-r-1)...1r(r-1)(r-2)...1}$$

Ejemplo:
De cuantas maneras se puede escoger un comité de 4 hombres de un grupo de 8?

$$C_4^8 = \frac{8!}{(8-4)!4!} = \frac{8(7)(6)(5)4!}{4!(4)(3)(2)1} = \frac{1680}{24} = 70$$

10.1.6 Permutaciones con Repetición

En el caso de las permutaciones, si el elmento1 se repite r_1 veces, el elemento 2 se repite r_2 veces, etc. Y el elemento *k* se repite r_k, se le llama permutaciones con repetición y se calcula con:

$$P_{(r_1,r_2,...r_k)}^n = \frac{n!}{r_1!r_2!...r_k!}$$

Ejemplo:
Cuantas palabras diferentes, aun sin significado, se pueden formar con las letras de la palabra **amorosos**?

$$P_{(1,1,3,1,2)}^8 = \frac{8!}{1(1)(3!)(1)2!} = \frac{8(7)(6)(5)(4)3!}{(3!)2!} = 3360$$

10.1.7 Variaciones con Repetición

En el caso de las variaciones si los elementos se pueden repetir hasta *r* veces se les denomina variaciones con repetición y se obtienen por:

$$VR_r^n = n^r$$

Ejemplo:
¿Cuantos números de cuatro cifras existen?

$$VR_4^{10} = 10^4 = 10000$$

EJERCICIOS PROPUESTOS

1. ¿De cuántas maneras se pueden colocar dos anillos diferentes en la misma mano, de modo que no estén en el mismo dedo?

2. Al lanzar cinco dados de distintos colores ¿cuántos resultados podemos obtener?

3. Con los números 1,2,3,4,5 y 6:

3.1 ¿Cuántos números distintos de siete cifras podríamos formar?
3.2 ¿Podremos numerar a los 3224564 habitantes de una ciudad con esos números?

4. Se lanzan al aire uno tras otro cinco dados equilibrados de seis caras. ¿Cuál es el número de casos posibles?

5. ¿Cuántos números de seis cifras existen que estén formados por cuatro números dos y por dos números tres?

6. Lola tiene 25 bolitas (10 rojas, 8 azules y 7 blancas) para hacerse un collar. En- garzando las 25 bolitas en un hilo, ¿cuántos collares distintos podrá realizar?

7. ¿Cuántas palabras distintas, con o sin sentido, podremos formar con las letras de la palabra educación? ¿y con la palabra vacaciones?

8. Un grupo de amigos formado por Raúl, Sonia, Ricardo y Carmen organizan una fiesta, acuerdan que dos de ellos se encargarán de comprar la comida y las bebidas ¿De cuántas formas posibles puede estar compuesta la pareja encargada de dicha misión?

9. Una fábrica de helados dispone de cinco sabores distintos (vainilla, chocolate, nata, fresa y cola) y quiere hacer helados de dos sabores ¿Cuántos tipos de helado podrán fabricar?

10. Un grupo de amigos y amigas se encuentran y se dan un beso para saludarse. Si se han dado en total 21 besos, ¿cuántas personas había?

11. En una carrera de 500 metros participan doce corredores ¿De cuántas maneras pueden adjudicarse las medallas de oro, plata, bronce?

12. ¿De cuántas formas pueden cubrirse los cargos de presidente, vicepresidente, secretario y tesorero de un club deportivo sabiendo que hay 14 candidatos?

[1] A *n!* se le denomina factorial de n

10. Nociones de Probabilidad (Eventos)

10.2 DEFINICIÓN DE PROBABILIDAD

Los eventos aleatorios no son predecibles con absoluta certeza, no obstante podemos medir el grado de confianza con que se hace un pronóstico, sobre la ocurrencia o no de un determinado suceso.

10.2.1 Probabilidad Clásica o "a priori"

Si un evento puede ocurrir de *n* maneras, equiprobables y mutuamente excluyentes, de las cuales *m* maneras son favorables al suceso *A*; se define probabilidad del suceso *A* como:

$$p(A) = \frac{m}{n} \quad \rightarrow$$

$$p(A) = \frac{casos\ favorables\ \ al\ evento\ \ A}{total\ de\ casos\ posibles\ \ del\ \exp erimento}$$

Ejemplo:
En el lanzamiento de un dado de seis caras una vez, si

$$A: \{obtener\ \ un\ número\ impar\} = \{1,3,5\}$$

$$S: \{1,2,3,4,5,6\}$$

$$p(A) = \frac{m}{n} \quad \rightarrow$$

$$p(A) = \frac{casos\ favorables\ \ al\ evento\ \ A}{total\ de\ casos\ posibles\ \ del\ \exp erimento} = \frac{3}{6}$$

10.2.2 Probabilidad "a posteriori" o de Frecuencia Relativa

Si un experimento se repite *n* veces $(n \rightarrow \infty)$, de las cuales *m* veces se presenta el suceso *A*, entonces es de esperarse que:

$$p(A) = \lim_{n \to \infty} \left(\frac{m}{n} \right) = p$$

La proporción de veces que se presenta el suceso *A* tiende a estabilizarse en un número entre 0 y 1 llamado probabilidad de *A.*

Si por ejemplo, lanzamos un dado cien veces y observamos la presencia del número "2" en 16 veces,

en tal caso

$$p(A) = \frac{16}{100}$$

10.2.3 Probabilidad Subjetiva

En la probabilidad subjetiva intervienen preferencias y emociones del analista que en general, son diferentes para cada caso. Por ejemplo, un apostador puede preferir el número "*3*" porque su horóscopo se lo recomienda.

10.3 AXIOMAS DE LA TEORÍA DE PROBABILIDADES

Para todo experimento, la probabilidad de ocurrencia de un evento *A, p(A),* es una función que cumple con los siguientes axiomas:

10.3.1 $p(A) \geq 0$ *toda probabilidad es no negativa*

10.3.2 $p(S) = 1$ *la probabilidad del espacio muestral es* 1

10.3.3 Si dos o más sucesos son incompatibles entre sí, entonces la probabilidad de la unión de ellos, es igual a la suma de sus probabilidades respectivas

$$si \quad p(A \cap B) = \Phi \quad \rightarrow \quad p(A \cup B) = p(A) + p(B)$$

De estos tres axiomas podemos, fácilmente, deducir que:

10.3.3.1 $p(\Phi) = 0$ La Probabilidad de un evento imposible es igual a cero.

10.3.3.2 $p(A) = 1 - p(\overline{A})$ La probabilidad de un evento es igual a la unidad menos la probabilidad de su complemento.

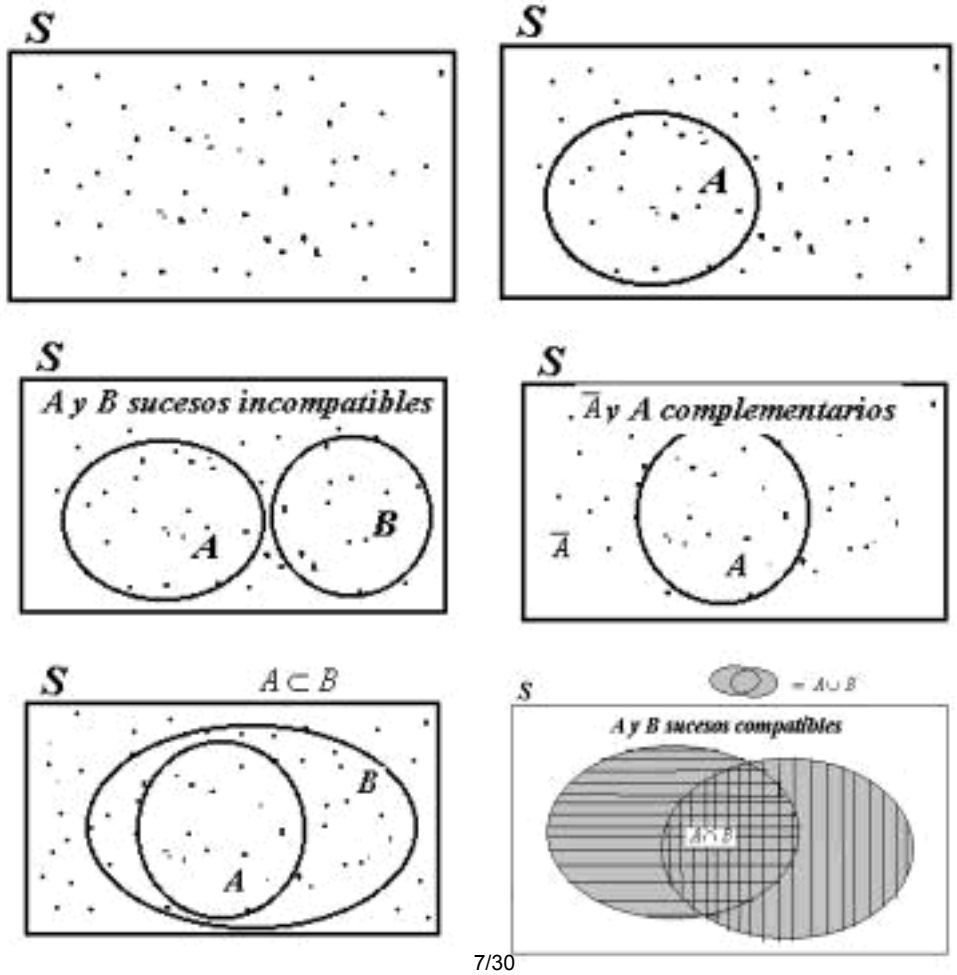
10.3.3.3 $0 \leq p(A) \leq 1$ Toda probabilidad está definida entre la probabilidad del suceso imposible y la probabilidad del evento seguro.

10.3.3.4 $Si \quad A \subset B \quad \rightarrow p(A) \leq p(B)$.

10.3.3.5 $Si \quad A \cap B \neq \Phi \quad \rightarrow$

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

Si dos eventos son compatibles, la probabilidad de su unión es igual a la suma de sus probabilidades menos la probabilidad de su intersección.



En el ejemplo del lanzamiento de dos dados si:

***A** : (suma sea mayor que 5 pero menor que 10)*

$$A=\left\{\begin{array}{l}(1,5)\ (1.6)\ (2,4)\ (2,5)\ (2,6)\ (3,3)\ (3,4)\\ (3,5)\ (3,6)\ (4,2)\ (4,3)\ (4,4)\ (4,5)\ (5,1)\\ (5,2)\ (5,3)\ (5,4)\ (6,1)\ (6,2)\ (6,3)\end{array}\right\}$$

$$p(A)=p\left\{\begin{array}{l}(1,5)\ (1.6)\ (2,4)\ (2,5)\ (2,6)\ (3,3)\ (3,4)\\ (3,5)\ (3,6)\ (4,2)\ (4,3)\ (4,4)\ (4,5)\ (5,1)\\ (5,2)\ (5,3)\ (5,4)\ (6,1)\ (6,2)\ (6,3)\end{array}\right\}=\frac{m}{n}=\frac{20}{36}$$

***B** : (la suma sea mayor que 8)*

$$B=\{(3,6)\ (4,5)\ (4,6)\ (5,4)\ (5,5)\ (5,6)\ (6,3)\ (6,4)\ (6,5)\ (6,6)\}$$

$$p(B)=p\left\{\begin{array}{l}(3,6)\ (4,5)\ (4,6)\ (5,4)\ (5,5)\\ (5,6)\ (6,3)\ (6,4)\ (6,5)\ (6,6)\end{array}\right\}=\frac{m}{n}=\frac{10}{36}$$

$$A\cap B=\{(3,6)\ (4,5)\ (5,4)\ (6,3)\}$$

$$p(A\cap B)=p\{(3,6)\ (4,5)\ (5,4)\ (6,3)\}=\frac{m}{n}=\frac{4}{36}$$

$$p(A\cup B)=p(A)+p(B)-p(A\cap B)\ \rightarrow$$

$$p(A\cup B)=\frac{20+10-4}{36}=\frac{26}{36}$$

10.4 PROBABILIDAD CONDICIONAL E INDEPENDENCIA ESTADÍSTICA

Si tenemos los sucesos *A, B* en un experimento aleatorio, con *p(B)>0*, se llama *probabilidad condicional a: p(A/B)* La probabilidad de ocurrencia del evento *"A"* dado que

ya se ha presentado el suceso *"B"*.

$$p(A/B) = \frac{p(A \cap B)}{p(B)} \quad , \quad p(B) > 0$$

Ejemplo:
a un grupo de personas se le pregunta sobre la intención de voto para las próximas elecciones.

Sexo / Intención	Masculino	Femenino	Total
Votará	140	80	220
No votará	40	60	100
Total	180	140	320

$$p(M) = \frac{180}{320} \qquad p(F) = \frac{140}{320}$$

$$p(V) = \frac{220}{320} \qquad p(nV) = \frac{100}{320}$$

p(vote dado que es masculino)=

$$p(V/M) = \frac{p(V \cap M)}{p(M)} = \frac{\frac{140}{320}}{\frac{180}{320}} = \frac{140}{180}$$

p(vote dado que es femenino)=

$$p(V/F) = \frac{p(V \cap F)}{p(F)} = \frac{80}{140}$$

Independencia Estadística

$$"A","B" \quad eventos \quad independientes \leftrightarrow$$
$$p(A \cap B) = p(A).p(B)$$

Por ejemplo la probabilidad de obtener un número impar en el segundo lanzamiento de un dado, no depende de si en el primer lanzamiento se

obtuvo un número impar.

CUESTIONARIO Y EJERCICIOS PROPUESTOS

1. Defina:

1.1 Experimento aleatorio, y experimento determinístico
1.2 Evento elemental, suceso seguro, suceso imposible, eventos excluyentes y eventos independientes.

2. Para cada uno de los eventos definidos en el lanzamiento de dos dados, calcular su respectiva probabilidad de ocurrencia.

3. En el experimento de seleccionar una ficha de dominó, determinar las probabilidades para todos sus eventos elementales.

4. Para el ejemplo de la intención de voto según el sexo, calcular la probabilidad de no votante dado que es de sexo masculino.

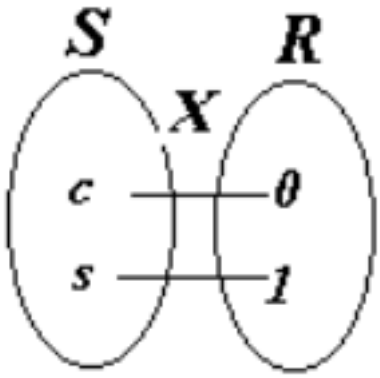
10. Nociones de Probabilidad (Eventos)

10.5 VARIABLE ALEATORIA

En el cálculo de probabilidades, generalmente, es más sencillo identificar los eventos numéricamente, y no con la simple descripción del suceso que pueda ocurrir, es más, en muchas ocasiones no podemos registrar todos los sucesos inmersos en el espacio muestral del experimento. Debemos recurrir a cuantificar esos símbolos iniciales en números reales que se puedan operar matemáticamente.

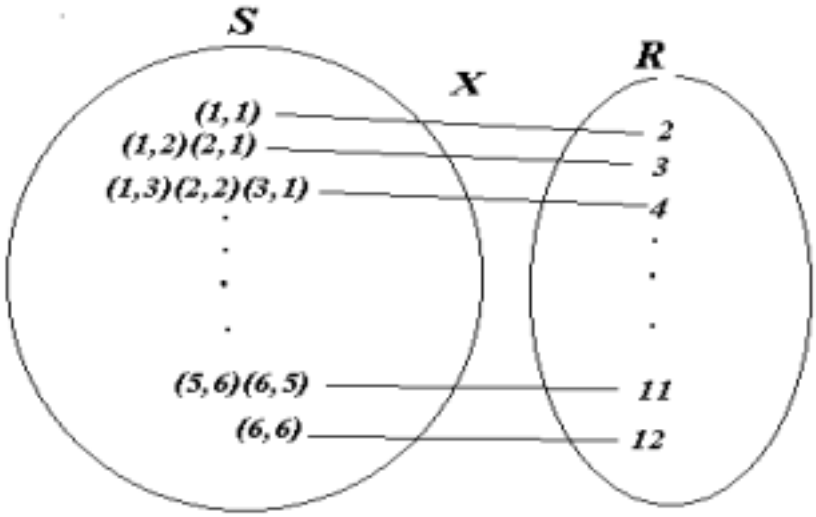
Definición: Una variable aleatoria es una función definida sobre un espacio muestral a los números reales. Si ese espacio muestral especificado como dominio es numerable, decimos que la variable es de tipo *discreto*, en caso contrario diremos que es de tipo *continuo.*

En el experimento de lanzar una moneda, una vez, definimos la variable aleatoria X: el número de sellos obtenido.



X(c) = 0
X(s) = 1

En la tirada de dos dados si X es la suma obtenida:



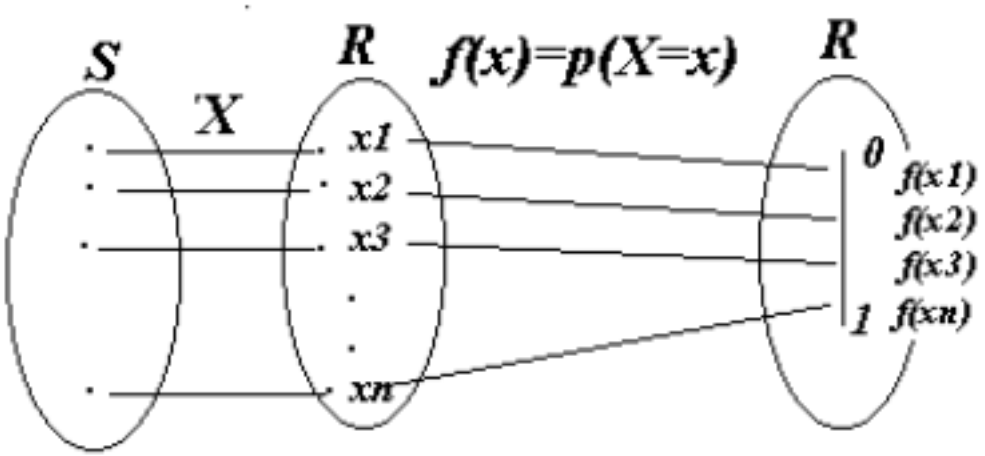
$$X(1,1) = 2, \quad X(1,2) = X(2,1) = 3, \quad X(1,3) = X(2,2) = X(3,1) = 4, \quad \cdots X(5,6) = X(6,5) = 11, \quad X(6,6) = 12$$

10.6 FUNCIÓN DE PROBABILIDAD

Las variables aleatorias, transforman eventos del espacio muestral en eventos numéricos, los cuales desde luego, tienen asociada una probabilidad de ocurrencia.

10.6.1 Función de Probabilidad *f(x)=p(X=x) :*

Es una función definida sobre una variable aleatoria a los reales en el intervalo $[0,1]$ que cumple con los axiomas de la teoría de la probabilidad.



10.6.2 Función de Distribución *F(x)=p(X=x)*

Es la acumulada de una función de probabilidad.

$$F(x) = p(X \le x_k) = \sum_{1}^{k} f(x_i) \qquad si \quad X \quad es \quad discreta$$

$$F(x) = p(X \le x_k) = \int_{-\infty}^{x_1} f(x)dx \qquad si \quad X \quad es \quad continua$$

-∞ : Limite inferior de la variable X

Ejemplo:

En el Lanzamiento de una Moneda,
X: Número de Sellos

X	**0**	**1**
f(x) = p(X=x)	$\frac{1}{2}$	$\frac{1}{2}$
F(x)= p(X≤x)	$\frac{1}{2}$	1

$$f(x) = p(X = x) = \begin{cases} \frac{1}{2} & para \quad x = 0,1 \\ 0 & en \quad otro \quad caso \end{cases}$$

Ejemplo:

X es la Suma Obtenida en el Lanzamiento de dos Dados:

X	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
$f(x) = p(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
$F(x) = p(X \le x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	1

. . . . .

$$f(x)=\begin{cases}\dfrac{x-1}{36} & para \quad x=2,3,4,\cdots 7\\ \dfrac{13-x}{36} & para \quad x=8,9,\cdots 12\\ 0 & en\ otro\ caso\end{cases}$$

Ejemplo:

Si X: Diferencia en Valor Absoluto,
Entre los dos Sectores de una Ficha de Dominó:

X	0	1	2	3	4	5	6
$f(x)=$ $p(X=x)$	$\frac{7}{28}$	$\frac{6}{28}$	$\frac{5}{28}$	$\frac{4}{28}$	$\frac{3}{28}$	$\frac{2}{28}$	$\frac{1}{28}$
$F(x)=$ $p(X\le x)$	$\frac{7}{28}$	$\frac{13}{28}$	$\frac{18}{28}$	$\frac{22}{28}$	$\frac{25}{28}$	$\frac{27}{28}$	1

$$f(x)=p(X=x)=\begin{cases}\dfrac{7-x}{28} & para \quad x=0,1,2,\cdots 6\\ 0 & en\ \ otro\ \ caso\end{cases}$$

Hemos creado 3 ejemplos de funciones de probabilidad para variables aleatorias discretas con sus respectivas funciones de distribución, que nos permiten calcular las probabilidades para cualquier tipo de evento. Calculemos algunas para el lanzamiento del par de dados, donde X es la suma obtenida:

$$A=\left\{x/X\le 9\right\} \qquad B=\left\{x/X>4\right\}$$
$$C=\left\{x/X<4\right\} \qquad D=\left\{x/X>10\right\}$$

Consultando directamente en la función de distribución de esta variable discreta, F(x)=p(X≤x) tenemos:

$$p(A) = p\left\{x/X \le 9\right\} = p(X \le 9) = \frac{30}{36} = \frac{5}{6}$$

$$p(B) = p\left\{x/X > 4\right\} = 1 - p(\overline{B}) = 1 - p\left\{x/X \le 4\right\}$$

$$= 1 - p(X \le 4) = 1 - \frac{6}{36} = \frac{30}{36} = \frac{5}{6}$$

$$p(C) = p\left\{x/X < 4\right\} = p(X < 4) = p(X \le 3) = \frac{3}{36} = \frac{1}{12}$$

$$p(D) = p\left\{x/X > 10\right\} = p(X > 10) = 1 - p(\overline{D}) =$$

$$1 - p(X \le 10) = 1 - \frac{33}{36} = \frac{3}{36} = \frac{1}{12}$$

$$p(A \cap B) = p\left[\left\{x/X \le 9\right\} \cap \left\{x/X > 4\right\}\right] = p(4 < X \le 9)$$

$$= p(X \le 9) - p(X \le 4) = \frac{30}{36} - \frac{6}{36} = \frac{24}{36} = \frac{2}{3}$$

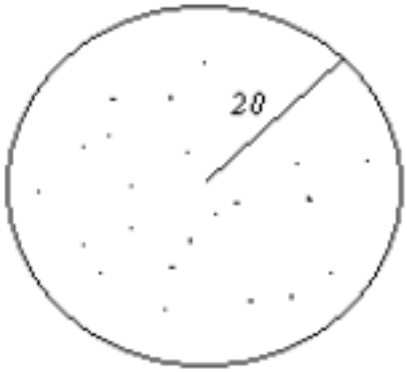
$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

$$= \frac{30}{36} + \frac{30}{36} - \frac{24}{36} = \frac{36}{36} = 1$$

$$p(C \cap D) = p\left[\left\{x/X < 4\right\} \cap \left\{x/X > 10\right\}\right] = p(\Phi) = 0$$

$$p(C \cup D) = p(C) + p(D) = \frac{3}{36} + \frac{3}{36} = \frac{6}{36} = \frac{1}{6}$$

Para el caso continuo, supongamos que un practicante de tiro al blanco siempre acierta indistintamente, en un círculo de 20 centímetros de radio.



La distancia que hay entre el punto “a=0” (centro) y cualquier punto de la circunferencia “b=20” es $b-a=20-0=20$.

$$f(x)=\begin{cases}\dfrac{1}{b-a}=\dfrac{1}{20} & para \quad 0\le x\le 20\\ 0 & en \quad otro \quad caso\end{cases}$$

¿ Cuál es la probabilidad que un disparo impacte a menos de 15 cm del centro? ¿ a más de 9 centímetros? ¿Entre 7 y 14 centímetros?

Para toda variable continua:

$$p(X=x)=0 \quad \rightarrow \quad p(X\le x)=p(X<x)$$

$$p(X<15)=\int_0^{15}\frac{1}{20}dx=\frac{x}{20}\Big]_0^{15} \quad =\frac{15}{20}-0=\frac{15}{20}$$

$$p(X>9)=1-p(X\le 9)=1-\int_0^{9}\frac{1}{20}dx=1-\frac{x}{20}\Big]_0^{9}$$

$$=1-\frac{9}{20}=\frac{11}{20}$$

$$p(7<X\le 14)=\int_7^{14}\frac{1}{20}dx=\frac{x}{20}\Big]_7^{14} \quad =\frac{14}{20}-\frac{7}{20}=\frac{7}{20}$$

CUESTIONARIO Y EJERCICIOS PROPUESTOS

1. Defina: Variable aleatoria, variable aleatoria discreta, variable aleatoria continua, función de probabilidad y función de distribución.

2. En el ejercicio de la ficha de dominó, si X representa la diferencia absoluta entre los dos números, representar y calcular la probabilidad de ocurrencia de los siguientes eventos:

2.1 La diferencia sea menor o igual a 5
2.2 La diferencia sea mayor que 2
2.3 La diferencia sea mayor que 2 pero menor o igual 5
2.4 La diferencia sea mayor que 5 ó menor que 3

10. Nociones de Probabilidad (Eventos)

10.7 VALOR ESPERADO (ESPERANZA MATEMÁTICA)

10.7.1 Media Aritmética Poblacional

En el tratamiento de las medidas de tendencia central, resaltamos la importancia de la media aritmética de una variable, como parámetro representativo de una muestra.

En el análisis poblacional, la media aritmética o *valor esperado* de una variable aleatoria, se define como el promedio ponderado de los diferentes valores que puede asumir la variable X, usando como ponderaciones las probabilidades respectivas de ocurrencia.

$$E(X) = \mu_X = x_1 f(x_1) + x_2 f(x_2) + \cdots + x_N f(x_N) = \sum_{1}^{N} x_i f(x_i)$$

si X es discreta ó

$$E(X) = \mu_X = \int_{-\infty}^{\infty} x f(x) dx$$

si X es continua

-∞ : límite inferior de la variable

∞ : límite superior de la variable

Ejemplo:

X es la Suma Obtenida en el Lanzamiento de Dos Dados

X	2	3	4	5	6	7	8	9	10	11	12
$f(x) = p(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
$xf(x)$	$\frac{2}{36}$	$\frac{6}{36}$	$\frac{12}{36}$	$\frac{20}{36}$	$\frac{30}{36}$	$\frac{42}{36}$	$\frac{40}{36}$	$\frac{36}{36}$	$\frac{30}{36}$	$\frac{22}{36}$	$\frac{12}{36}$

$$E(X) = \mu_x = \sum_1^N xf(x) = \frac{252}{36} = 7$$

En promedio la suma obtenida en N tiradas es de “7”. Si pagaramos en pesos la suma obtenida en cada lanzamiento, deberíamos cobrar más de 7 pesos para obtener utilidad en el juego.

En la variable X, distancia del centro al punto de impacto del tirador, el valor esperado es:

$$E(x) = \mu_x = \int_0^{20} x\frac{1}{20}dx = \frac{x^2}{40}\Big]_0^{20} = \frac{400}{40} = 10$$

10.7.2 Varianza Poblacional

Similarmente a la definición de la media aritmética poblacional, la varianza se define como:

$$\sigma_x^2 = E\left[(X-\mu)^2\right] = \sum_1^N (x_i - \mu)^2 f(x_i) \quad si\ X\ es\ discreta$$

ó

$$\sigma_x^2 = E\left[(X-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \quad si\ X\ es\ continua$$

CUESTIONARIO Y EJERCICIOS PROPUESTOS

1. Calcular el valor esperado para la variable diferencia en el ejemplo del dominó.

2. Si usted juega chance, calcule su valor real de acuerdo con los premios que espera obtener y compárelo con lo que realmente paga.

3. Tome un billete de lotería y calcule su precio equitativo.

4. Un contrabandista se enfrenta al siguiente dilema: Introducir o no, mercancía por valor de 5'000.000 obteniendo una utilidad de 1'000.000. El riesgo de ser detectado y castigado con el decomiso de la mercancía es del 17%. ¿Que le aconseja usted.?

11. Distribuciones Especiales

En el capítulo anterior desarrollamos modelos probabilísticos a partir de abstracciones de los experimentos previamente descritos, a los cuales se les crea una función de probabilidad, que describa las posibilidades de esa realidad experimental.

Muchos de los acontecimientos cotidianos, pueden ser asimilados a funciones probabilísticas teóricas, que son de gran ayuda en la toma de decisiones bajo condiciones de incertidumbre. Eminentes estudiosos de la estadística han planteado modelos probabilísticos que han contribuido al desarrollo de la ciencia. Veamos algunos de ellos:

11.1 DISTRIBUCIÓN DE BERNOULLI

Se puede afirmar que el experimento de Bernoulli, describe el modelo aleatorio más sencillo, el cual tiene las siguientes características:

- En el experimento sólo se hace un ensayo.
- En el experimento sólo se admiten dos resultados incompatibles, que llamaremos éxito y fracaso.
- La probabilidad de un éxito es $p(E)=p$.
- La probabilidad de un fracaso es $p(F)=1-p = q$
- X : es el número de éxitos $x = 0,1$.

X	0	1
P(X=x) = f(x)	1-p	P
P(X≤x)=F(x)	1-p	1

$$f(x)=\begin{cases}1-p & si \quad x=0\\ p & si \quad x=1\\ 0 & en \quad otro \quad caso\end{cases}$$

Es el caso cuando se lanza una moneda una vez y se observa de cual lado cae o se analiza un artículo para ver si está defectuoso o no, se obtiene o no un trabajo etc.

11.2 DISTRIBUCIÓN BINOMIAL

La distribución binomial se obtiene haciendo *n* pruebas de Bernoulli independientes entre sí, en tal caso tiene las siguientes características:

- *n :* número de repeticiones independientes del experimento de Bernoulli.
- Todas las pruebas deben tener una probabilidad constante de éxito "*p*" y una probabilidad constante de fracaso "*q*"*=1-p*.
- X : es el número de éxitos en las *n* pruebas, entonces; *n-X* : número de fracasos.

Analicemos el experimento con tres repeticiones:

$$n=3, \qquad p(E)=p, \qquad p(F)=1-p$$
$$X: número \quad de \quad éxitos \quad x=0,1,2,3$$

$$para \quad X=0 \rightarrow FFF \quad por\ la\ independencia\ se\ tiene$$

$$p(X=0)=(1-p)(1-p)(1-p)=(1-p)^3. \quad (1)$$

$$para \quad X=1 \rightarrow EFF \quad ó \quad FEF \quad ó \quad FFE$$
$$p(X=1)=p(1-p)(1-p)+(1-p)p(1-p)+(1-p)(1-p)p$$
$$=3p(1-p)^2 \qquad (2)$$

$$para \quad X=2 \rightarrow EEF \quad ó \quad EFE \quad ó \quad FEE$$
$$p(X=2)=pp(1-p)+p(1-p)p+(1-p)pp$$
$$=3p^2(1-p) \qquad (3)$$

$$para \quad X=3 \rightarrow EEE$$
$$p(X=3)=ppp=p^3 \qquad (4)$$

(1) se puede expresar como:

$$\frac{3!}{(3-0)!0!}p^0(1-p)^{3-0} = C_0^3 p^0(1-p)^{3-0}$$

(2) se puede expresar como:

$$\frac{3!}{(3-1)!1!}p^1(1-p)^{3-1} = C_1^3 p^1(1-p)^{3-1}$$

(3) se puede expresar como:

$$\frac{3!}{(3-2)!2!}p^2(1-p)^{3-2} = C_2^3 p^2(1-p)^{3-2}$$

(4) se puede expresar como:

$$\frac{3!}{(3-3)!3!}p^3(1-p)^{3-3} = C_3^3 p^3(1-p)^{3-3}$$

entonces para *n=3*, tenemos que:

$$f(x) = \frac{3!}{(3-x)!\,x!}p^x(1-p)^{3-x} = C_x^3 p^x(1-p)^{3-x} \quad \rightarrow$$

$$f(x) = C_x^3 p^x(1-p)^{3-x}, \quad para \quad x = 0,1,2,3$$
$$= 0 \quad en \quad otro \quad caso$$

$$F(x) = p(X \le x) = \sum_0^x C_x^3 p^x(1-p)^{3-x}$$

X	0	1	2	3
$f(x) = C_x^3 p^x (1-p)^{3-x}$	$(1-p)^3$	$3p^1(1-p)^2$	$3p^2(1-p)^1$	p^3
$F(x) = \sum_0^x C_x^3 p^x (1-p)^{3-x}$	$(1-p)^3$	$(1-p)^3 + 3p^1(1-p)^2$	$(1-p)^3 3p^1(1-p)^2 + 3p^2(1-p)^1$	1

En general la función de probabilidad binomial tiene la siguiente forma:

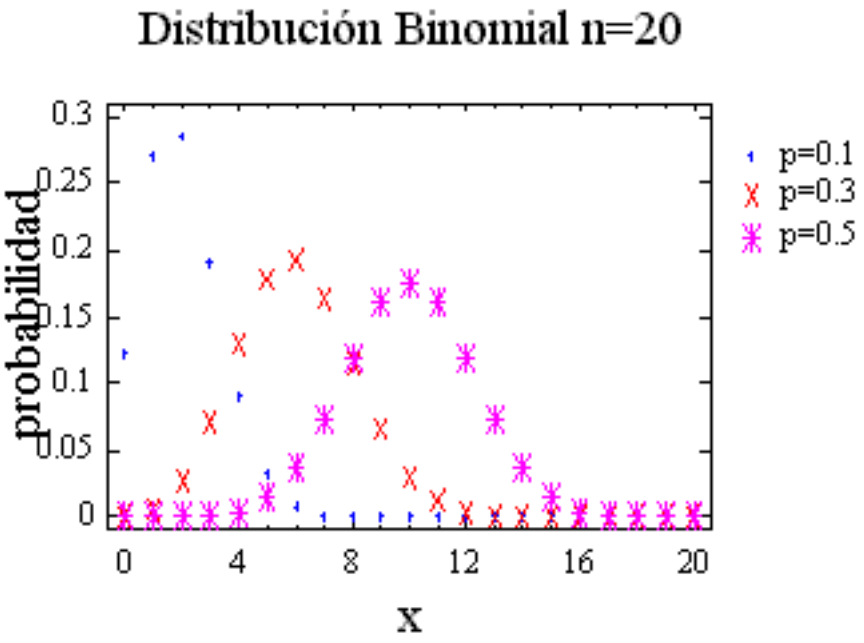
$$f(x) = C_x^n p^x (1-p)^{n-x} \quad para \quad x = 0,1,2,3,\cdots,n$$
$$= 0 \quad en \quad otro \quad caso$$

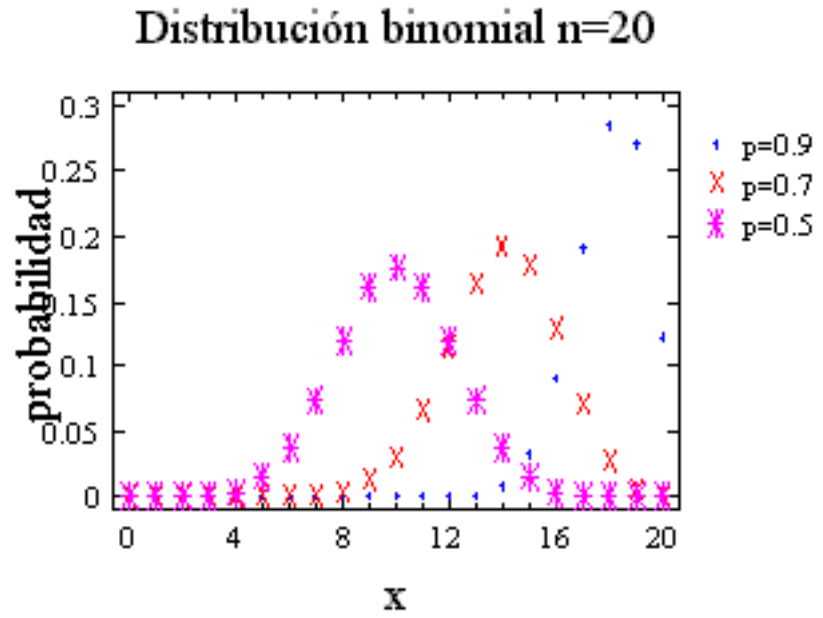
y la función de distribución:

$$F(x) = p(X \le x) = \sum_0^x C_x^n p^x (1-p)^{n-x}$$

La media aritmética de una variable aleatoria con distribución binomial es $E(X) = \mu_x = np$, y varianza $\sigma_x^2 = np(1-p)$. Con los parámetros *n*, y *p* se tipifica la distribución binomial y la representamos como: $X \sim b(n,p)$.

La distribución binomial es simétrica cuando *p=0.5*, en caso contrario es asimétrica a la izquierda o a la derecha, según el valor de *p* sea inferior o superior a 0.5. Ver gráfico:





Tablas Binomiales

$$F(x) = p(X \leq x) = \sum_{0}^{x} C_x^n p^x (1-p)^{n-x}$$

Probabilidades binomiales acumuladas *F(x), para "n=5"*									
	P								
X	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.590	0.328	0.168	0.078	0.031	0.010	0.002	0.000	0.000
1	0.919	0.737	0.528	0.337	0.188	0.087	0.031	0.007	0.000
2	0.991	0.942	0.837	0.683	0.500	0.317	0.163	0.058	0.009
3	1.000	0.993	0.969	0.913	0.813	0.663	0.472	0.263	0.081
4	1.000	1.000	0.998	0.990	0.969	0.922	0.832	0.672	0.410
5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

$$F(x) = p(X \le x) = \sum_{0}^{x} C_x^n p^x (1-p)^{n-x}$$

Probabilidades binomiales acumuladas *F(x), para "n=10"*									
	p								
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.349	0.107	0.028	0.006	0.001	0.000	0.000	0.000	0.000
1	0.736	0.376	0.149	0.046	0.011	0.002	0.000	0.000	0.000
2	0.930	0.678	0.383	0.167	0.055	0.012	0.002	0.000	0.000
3	0.987	0.879	0.650	0.382	0.172	0.055	0.011	0.001	0.000
4	0.998	0.967	0.850	0.633	0.377	0.166	0.047	0.006	0.000
5	1.000	0.994	0.953	0.834	0.623	0.367	0.150	0.033	0.002
6	1.000	0.999	0.989	0.945	0.828	0.618	0.350	0.121	0.013
7	1.000	1.000	0.998	0.988	0.945	0.833	0.617	0.322	0.070
8	1.000	1.000	1.000	0.998	0.989	0.954	0.851	0.624	0.264
9	1.000	1.000	1.000	1.000	0.999	0.994	0.972	0.893	0.651
10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

$$F(x) = p(X \le x) = \sum_{0}^{x} C_x^n p^x (1-p)^{n-x}$$

Probabilidades binomiales acumuladas *F(x), para "n=15"*									
	p								
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.206	0.035	0.005	0.000	0.000	0.000	0.000	0.000	0.000
1	0.549	0.167	0.035	0.005	0.000	0.000	0.000	0.000	0.000
2	0.816	0.398	0.127	0.027	0.004	0.000	0.000	0.000	0.000
3	0.944	0.648	0.297	0.091	0.018	0.002	0.000	0.000	0.000
4	0.987	0.836	0.515	0.217	0.059	0.009	0.001	0.000	0.000
5	0.998	0.939	0.722	0.403	0.151	0.034	0.004	0.000	0.000
6	1.000	0.982	0.869	0.610	0.304	0.095	0.015	0.001	0.000
7	1.000	0.996	0.950	0.787	0.500	0.213	0.050	0.004	0.000
8	1.000	0.999	0.985	0.905	0.696	0.390	0.131	0.018	0.000
9	1.000	1.000	0.996	0.966	0.849	0.597	0.278	0.061	0.002
10	1.000	1.000	0.999	0.991	0.941	0.783	0.485	0.164	0.013
11	1.000	1.000	1.000	0.998	0.982	0.909	0.703	0.352	0.056
12	1.000	1.000	1.000	1.000	0.996	0.973	0.873	0.602	0.184
13	1.000	1.000	1.000	1.000	1.000	0.995	0.965	0.833	0.451
14	1.000	1.000	1.000	1.000	1.000	1.000	0.995	0.965	0.794
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Probabilidades binomiales acumuladas *F(x), para "n=20"*									
	P								
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.122	0.012	0.001	0.000	0.000	0.000	0.000	0.000	0.000
1	0.392	0.069	0.008	0.001	0.000	0.000	0.000	0.000	0.000
2	0.677	0.206	0.035	0.004	0.000	0.000	0.000	0.000	0.000
3	0.867	0.411	0.107	0.016	0.001	0.000	0.000	0.000	0.000
4	0.957	0.630	0.238	0.051	0.006	0.000	0.000	0.000	0.000
5	0.989	0.804	0.416	0.126	0.021	0.002	0.000	0.000	0.000
6	0.998	0.913	0.608	0.250	0.058	0.006	0.000	0.000	0.000
7	1.000	0.968	0.772	0.416	0.132	0.021	0.001	0.000	0.000
8	1.000	0.990	0.887	0.596	0.252	0.057	0.005	0.000	0.000
9	1.000	0.997	0.952	0.755	0.412	0.128	0.017	0.001	0.000
10	1.000	0.999	0.983	0.872	0.588	0.245	0.048	0.003	0.000
11	1.000	1.000	0.995	0.943	0.748	0.404	0.113	0.010	0.000
12	1.000	1.000	0.999	0.979	0.868	0.584	0.228	0.032	0.000
13	1.000	1.000	1.000	0.994	0.942	0.750	0.392	0.087	0.002
14	1.000	1.000	1.000	0.998	0.979	0.874	0.584	0.196	0.011
15	1.000	1.000	1.000	1.000	0.994	0.949	0.762	0.370	0.043
16	1.000	1.000	1.000	1.000	0.999	0.984	0.893	0.589	0.133
17	1.000	1.000	1.000	1.000	1.000	0.996	0.965	0.794	0.323
18	1.000	1.000	1.000	1.000	1.000	0.999	0.992	0.931	0.608
19	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.988	0.878
20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

$$F(x) = p(X \le x) = \sum_{0}^{x} C_x^n p^x (1-p)^{n-x}$$

Probabilidades binomiales acumuladas *F(x), para "n=25"*									
	P								
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.072	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.271	0.027	0.002	0.000	0.000	0.000	0.000	0.000	0.000
2	0.537	0.098	0.009	0.000	0.000	0.000	0.000	0.000	0.000
3	0.764	0.234	0.033	0.002	0.000	0.000	0.000	0.000	0.000
4	0.902	0.421	0.090	0.009	0.000	0.000	0.000	0.000	0.000
5	0.967	0.617	0.193	0.029	0.002	0.000	0.000	0.000	0.000
6	0.991	0.780	0.341	0.074	0.007	0.000	0.000	0.000	0.000
7	0.998	0.891	0.512	0.154	0.022	0.001	0.000	0.000	0.000
8	1.000	0.953	0.677	0.274	0.054	0.004	0.000	0.000	0.000
9	1.000	0.983	0.811	0.425	0.115	0.013	0.000	0.000	0.000
10	1.000	0.994	0.902	0.586	0.212	0.034	0.002	0.000	0.000
11	1.000	0.998	0.956	0.732	0.345	0.078	0.006	0.000	0.000
12	1.000	1.000	0.983	0.846	0.500	0.154	0.017	0.000	0.000
13	1.000	1.000	0.994	0.922	0.655	0.268	0.044	0.002	0.000
14	1.000	1.000	0.998	0.966	0.788	0.414	0.098	0.006	0.000
15	1.000	1.000	1.000	0.987	0.885	0.575	0.189	0.017	0.000
16	1.000	1.000	1.000	0.996	0.946	0.726	0.323	0.047	0.000
17	1.000	1.000	1.000	0.999	0.978	0.846	0.488	0.109	0.002
18	1.000	1.000	1.000	1.000	0.993	0.926	0.659	0.220	0.009
19	1.000	1.000	1.000	1.000	0.998	0.971	0.807	0.383	0.033
20	1.000	1.000	1.000	1.000	1.000	0.991	0.910	0.579	0.098
21	1.000	1.000	1.000	1.000	1.000	0.998	0.967	0.766	0.236
22	1.000	1.000	1.000	1.000	1.000	1.000	0.991	0.902	0.463
23	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.973	0.729
24	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.928
25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

$$F(x) = p(X \le x) = \sum_{0}^{x} C_x^n p^x (1-p)^{n-x}$$

Probabilidades binomiales acumuladas *F(x), para "n=30"*									
	P								
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.042	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.184	0.011	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.411	0.044	0.002	0.000	0.000	0.000	0.000	0.000	0.000
3	0.647	0.123	0.009	0.000	0.000	0.000	0.000	0.000	0.000
4	0.825	0.255	0.030	0.002	0.000	0.000	0.000	0.000	0.000
5	0.927	0.428	0.077	0.006	0.000	0.000	0.000	0.000	0.000
6	0.974	0.607	0.160	0.017	0.001	0.000	0.000	0.000	0.000
7	0.992	0.761	0.281	0.044	0.003	0.000	0.000	0.000	0.000
8	0.998	0.871	0.432	0.094	0.008	0.000	0.000	0.000	0.000
9	1.000	0.939	0.589	0.176	0.021	0.001	0.000	0.000	0.000
10	1.000	0.974	0.730	0.291	0.049	0.003	0.000	0.000	0.000
11	1.000	0.991	0.841	0.431	0.100	0.008	0.000	0.000	0.000
12	1.000	0.997	0.916	0.578	0.181	0.021	0.001	0.000	0.000
13	1.000	0.999	0.960	0.715	0.292	0.048	0.002	0.000	0.000
14	1.000	1.000	0.983	0.825	0.428	0.097	0.006	0.000	0.000
15	1.000	1.000	0.994	0.903	0.572	0.175	0.017	0.000	0.000
16	1.000	1.000	0.998	0.952	0.708	0.285	0.040	0.001	0.000
17	1.000	1.000	0.999	0.979	0.819	0.422	0.084	0.003	0.000
18	1.000	1.000	1.000	0.992	0.900	0.569	0.159	0.009	0.000
19	1.000	1.000	1.000	0.997	0.951	0.709	0.270	0.026	0.000
20	1.000	1.000	1.000	0.999	0.979	0.824	0.411	0.061	0.000
21	1.000	1.000	1.000	1.000	0.992	0.906	0.568	0.129	0.002
22	1.000	1.000	1.000	1.000	0.997	0.956	0.719	0.239	0.008
23	1.000	1.000	1.000	1.000	0.999	0.983	0.840	0.393	0.026
24	1.000	1.000	1.000	1.000	1.000	0.994	0.923	0.572	0.073
25	1.000	1.000	1.000	1.000	1.000	0.998	0.970	0.745	0.175
26	1.000	1.000	1.000	1.000	1.000	1.000	0.991	0.877	0.353
27	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.956	0.589
28	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.989	0.816
29	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.958
30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Probabilidades binomiales acumuladas *F(x), para "n=40"*									
	P								
x	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0	0.015	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.080	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.223	0.008	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.423	0.028	0.001	0.000	0.000	0.000	0.000	0.000	0.000
4	0.629	0.076	0.003	0.000	0.000	0.000	0.000	0.000	0.000
5	0.794	0.161	0.009	0.000	0.000	0.000	0.000	0.000	0.000
6	0.900	0.286	0.024	0.001	0.000	0.000	0.000	0.000	0.000
7	0.958	0.437	0.055	0.002	0.000	0.000	0.000	0.000	0.000
8	0.985	0.593	0.111	0.006	0.000	0.000	0.000	0.000	0.000
9	0.995	0.732	0.196	0.016	0.000	0.000	0.000	0.000	0.000
10	0.999	0.839	0.309	0.035	0.001	0.000	0.000	0.000	0.000
11	1.000	0.912	0.441	0.071	0.003	0.000	0.000	0.000	0.000
12	1.000	0.957	0.577	0.129	0.008	0.000	0.000	0.000	0.000
13	1.000	0.981	0.703	0.211	0.019	0.000	0.000	0.000	0.000
14	1.000	0.992	0.807	0.317	0.040	0.001	0.000	0.000	0.000
15	1.000	0.997	0.885	0.440	0.077	0.003	0.000	0.000	0.000
16	1.000	0.999	0.937	0.568	0.134	0.008	0.000	0.000	0.000
17	1.000	1.000	0.968	0.689	0.215	0.019	0.000	0.000	0.000
18	1.000	1.000	0.985	0.791	0.318	0.039	0.001	0.000	0.000
19	1.000	1.000	0.994	0.870	0.437	0.074	0.002	0.000	0.000
20	1.000	1.000	0.998	0.926	0.563	0.130	0.006	0.000	0.000
21	1.000	1.000	0.999	0.961	0.682	0.209	0.015	0.000	0.000
22	1.000	1.000	1.000	0.981	0.785	0.311	0.032	0.000	0.000
23	1.000	1.000	1.000	0.992	0.866	0.432	0.063	0.001	0.000
24	1.000	1.000	1.000	0.997	0.923	0.560	0.115	0.003	0.000
25	1.000	1.000	1.000	0.999	0.960	0.683	0.193	0.008	0.000
26	1.000	1.000	1.000	1.000	0.981	0.789	0.297	0.019	0.000
27	1.000	1.000	1.000	1.000	0.992	0.871	0.423	0.043	0.000
28	1.000	1.000	1.000	1.000	0.997	0.929	0.559	0.088	0.000
29	1.000	1.000	1.000	1.000	0.999	0.965	0.691	0.161	0.001
30	1.000	1.000	1.000	1.000	1.000	0.984	0.804	0.268	0.005
31	1.000	1.000	1.000	1.000	1.000	0.994	0.889	0.407	0.015
32	1.000	1.000	1.000	1.000	1.000	0.998	0.945	0.563	0.042
33	1.000	1.000	1.000	1.000	1.000	0.999	0.976	0.714	0.100
34	1.000	1.000	1.000	1.000	1.000	1.000	0.991	0.839	0.206
35	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.924	0.371
36	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.972	0.577
37	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.992	0.777
38	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.920
39	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.985
40	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Probabilidades binomiales acumuladas F(x), para "n=50"									
	P								
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.034	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.112	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.250	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.431	0.018	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5	0.616	0.048	0.001	0.000	0.000	0.000	0.000	0.000	0.000
6	0.770	0.103	0.002	0.000	0.000	0.000	0.000	0.000	0.000
7	0.878	0.190	0.007	0.000	0.000	0.000	0.000	0.000	0.000
8	0.942	0.307	0.018	0.000	0.000	0.000	0.000	0.000	0.000
9	0.975	0.444	0.040	0.001	0.000	0.000	0.000	0.000	0.000
10	0.991	0.584	0.079	0.002	0.000	0.000	0.000	0.000	0.000
11	0.997	0.711	0.139	0.006	0.000	0.000	0.000	0.000	0.000
12	0.999	0.814	0.223	0.013	0.000	0.000	0.000	0.000	0.000
13	1.000	0.889	0.328	0.028	0.000	0.000	0.000	0.000	0.000
14	1.000	0.939	0.447	0.054	0.001	0.000	0.000	0.000	0.000
15	1.000	0.969	0.569	0.096	0.003	0.000	0.000	0.000	0.000
16	1.000	0.986	0.684	0.156	0.008	0.000	0.000	0.000	0.000
17	1.000	0.994	0.782	0.237	0.016	0.000	0.000	0.000	0.000
18	1.000	0.997	0.859	0.336	0.032	0.001	0.000	0.000	0.000
19	1.000	0.999	0.915	0.446	0.059	0.001	0.000	0.000	0.000
20	1.000	1.000	0.952	0.561	0.101	0.003	0.000	0.000	0.000
21	1.000	1.000	0.975	0.670	0.161	0.008	0.000	0.000	0.000
22	1.000	1.000	0.988	0.766	0.240	0.016	0.000	0.000	0.000
23	1.000	1.000	0.994	0.844	0.336	0.031	0.000	0.000	0.000
24	1.000	1.000	0.998	0.902	0.444	0.057	0.001	0.000	0.000
25	1.000	1.000	0.999	0.943	0.556	0.098	0.002	0.000	0.000
26	1.000	1.000	1.000	0.969	0.664	0.156	0.006	0.000	0.000
27	1.000	1.000	1.000	0.984	0.760	0.234	0.012	0.000	0.000
28	1.000	1.000	1.000	0.992	0.839	0.330	0.025	0.000	0.000
29	1.000	1.000	1.000	0.997	0.899	0.439	0.048	0.000	0.000
30	1.000	1.000	1.000	0.999	0.941	0.554	0.085	0.001	0.000

31	1.000	1.000	1.000	0.999	0.968	0.664	0.141	0.003	0.000
32	1.000	1.000	1.000	1.000	0.984	0.763	0.218	0.006	0.000
33	1.000	1.000	1.000	1.000	0.992	0.844	0.316	0.014	0.000
34	1.000	1.000	1.000	1.000	0.997	0.904	0.431	0.031	0.000
35	1.000	1.000	1.000	1.000	0.999	0.946	0.553	0.061	0.000
36	1.000	1.000	1.000	1.000	1.000	0.972	0.672	0.111	0.000
37	1.000	1.000	1.000	1.000	1.000	0.987	0.777	0.186	0.001
38	1.000	1.000	1.000	1.000	1.000	0.994	0.861	0.289	0.003
39	1.000	1.000	1.000	1.000	1.000	0.998	0.921	0.416	0.009
40	1.000	1.000	1.000	1.000	1.000	0.999	0.960	0.556	0.025
41	1.000	1.000	1.000	1.000	1.000	1.000	0.982	0.693	0.058
42	1.000	1.000	1.000	1.000	1.000	1.000	0.993	0.810	0.122
43	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.897	0.230
44	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.952	0.384
45	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.982	0.569
46	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.994	0.750
47	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.888
48	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.966
49	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995
50	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Ejemplo:

Se sabe que el 20% de la cartera de una empresa está vencida, se toma una muestra al azar de 15 cuentas. ¿Cuál es la probabilidad de que:

1. Haya cuatro ó menos cuentas vencidas?
2. Haya menos de cuatro cuentas vencidas?
3. Haya más de dos cuentas vencidas.
4. Haya más de dos pero menos de cinco cuentas vencidas?
5. Haya exactamente 3 cuentas vencidas?
6. No haya cuentas vencidas?
7. Cuál es valor esperado de cuentas vencidas?
8. Cuál es la desviación estándar para el número de cuentas vencidas?

Solución:

X: número de cuentas vencidas.
Éxito: Cuenta vencida.
Probabilidad de éxito : p=0.2
Número de pruebas n=15

1. En las tablas de distribuciones binomiales, $b \sim (15,0.2)$, en la intersección *x=4* y *p=0.2*, consultamos $p(X \leq 4) = 0.83577$.

2. $p(X < 4) = p(X \leq 3) = 0.64816$

3. $p(X > 2) = 1 - p(X \leq 2) = 1 - 0.39802 = 0.60198$

4. $p(2 < X < 5) = p(2 < X \leq 4) = p(X \leq 4) - p(X \leq 2)$
$= 0.83577 - 0.39802 = 0.43775$

5. $p(X = 3) = p(X \leq 3) - p(X \leq 2) =$
$0.64816 - 0.39802 = 0.25014$

6. $p(X = 0) = 0.03518$

7. $E(X) = np = 15(0.2) = 3$

8. $\sigma_x^2 = np(1-p) = 15(0.2)(0.8) = 2.4 \rightarrow \sigma_x$
$= \sqrt{\sigma_x^2} = \sqrt{2.4} = 1.5491$

11. Distribuciones Especiales

11.3 DISTRIBUCIÓN DE POISSON

La distribución de Poisson es de gran utilidad cuando tenemos variables distribuidas a través del tiempo ó del espacio. Es el caso del número de llamadas que entran a una central telefónica en una unidad de tiempo, la cantidad de personas que atiende un cajero en una hora, los baches por kilómetro en una autopista, los artículos defectuosos que hay en un lote de producción; amén de su utilización como aproximación binomial cuando *p* es muy cercano a cero, o n superior a 30. (*p<0.1 , n>30)*.

La función de probabilidad de Poisson es:

$$p(X=x)=f(x)=\begin{cases}\dfrac{\mu^x e^{-\mu}}{x!} & para \quad x=0,1,2,3,\cdots,\cdots \\ 0 & en \quad otro \quad caso\end{cases}$$

$$F(x)=p(X\le x)=\sum_0^x \frac{\mu^x e^{-\mu}}{x!}$$

Donde:

$\mu=E(X)=\sigma_x^2$: es decir, la media aritmética es igual a la varianza.

$e=2.71828$: (la base de los logaritmos naturales).

X : número de éxitos en la unidad de tiempo o de espacio considerado.

Ejemplo:

Un cajero de un banco atiende en promedio 7 personas por hora, cual es la probabilidad de que un una hora determinada:

1. Atienda menos de 5 personas
2. Atienda más de 8 personas
3. Atienda más de 5 pero menos de 8 personas

4. Atienda exactamente 7 personas

Consultando la tabla para la distribución de Poisson:

1. $p(X<5)=p(X\leq 4)=0.173$

2. $p(X>8)=1-p(X\leq 8)=1-0.729=0.271$

3. $p(5<X<8)=p(5<X\leq 7)$
$=p(X\leq 7)-p(X\leq 5)$
$=0.599-0.301=0.298$

4. $p(X=7)=p(X\leq 7)-p(X\leq 6)=$
$0.599-0.450=0.149$

Ejemplo:

En cierto núcleo poblacional, el 0.5% es portador del V.I.H. En una muestra de 80 personas, cual es la probabilidad:

1. De que haya alguna persona portadora.
2. No haya personas portadoras.

Solución:

$p=0.005 \qquad n=80 \qquad np=0.4$

1. $p(X>0)=1-p(X\leq 0)=1-0.67=0.33$
2. $p(X=0)=0.67$

Probabilidades de Poisson Acumuladas

$$F(x)=p(X\leq x)=\sum_{0}^{x}\frac{\mu^{x}e^{-\mu}}{x!}$$

	x											
u	0	1	2	3	4	5	6	7	8	9	10	11
0.2	0.819	0.982	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.4	0.670	0.938	0.992	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.6	0.549	0.878	0.977	0.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.8	0.449	0.809	0.953	0.991	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1	0.368	0.736	0.920	0.981	0.996	0.999	1.000	1.000	1.000	1.000	1.000	1.000
1.2	0.301	0.663	0.879	0.966	0.992	0.998	1.000	1.000	1.000	1.000	1.000	1.000
1.4	0.247	0.592	0.833	0.946	0.986	0.997	0.999	1.000	1.000	1.000	1.000	1.000
1.6	0.202	0.525	0.783	0.921	0.976	0.994	0.999	1.000	1.000	1.000	1.000	1.000
1.8	0.165	0.463	0.731	0.891	0.964	0.990	0.997	0.999	1.000	1.000	1.000	1.000
2	0.135	0.406	0.677	0.857	0.947	0.983	0.995	0.999	1.000	1.000	1.000	1.000
2.2	0.111	0.355	0.623	0.819	0.928	0.975	0.993	0.998	1.000	1.000	1.000	1.000
2.4	0.091	0.308	0.570	0.779	0.904	0.964	0.988	0.997	0.999	1.000	1.000	1.000
2.6	0.074	0.267	0.518	0.736	0.877	0.951	0.983	0.995	0.999	1.000	1.000	1.000
2.8	0.061	0.231	0.469	0.692	0.848	0.935	0.976	0.992	0.998	0.999	1.000	1.000
3	0.050	0.199	0.423	0.647	0.815	0.916	0.966	0.988	0.996	0.999	1.000	1.000
3.2	0.041	0.171	0.380	0.603	0.781	0.895	0.955	0.983	0.994	0.998	1.000	1.000
3.8	0.022	0.107	0.269	0.473	0.668	0.816	0.909	0.960	0.984	0.994	0.998	0.999
4	0.018	0.092	0.238	0.433	0.629	0.785	0.889	0.949	0.979	0.992	0.997	0.999
4.2	0.015	0.078	0.210	0.395	0.590	0.753	0.867	0.936	0.972	0.989	0.996	0.999
4.4	0.012	0.066	0.185	0.359	0.551	0.720	0.844	0.921	0.964	0.985	0.994	0.998
4.6	0.010	0.056	0.163	0.326	0.513	0.686	0.818	0.905	0.955	0.980	0.992	0.997
4.8	0.008	0.048	0.143	0.294	0.476	0.651	0.791	0.887	0.944	0.975	0.990	0.996
5	0.007	0.040	0.125	0.265	0.440	0.616	0.762	0.867	0.932	0.968	0.986	0.995
5.2	0.006	0.034	0.109	0.238	0.406	0.581	0.732	0.845	0.918	0.960	0.982	0.993
5.4	0.005	0.029	0.095	0.213	0.373	0.546	0.702	0.822	0.903	0.951	0.977	0.990

11.4 DISTRIBUCIÓN NORMAL

Dada la caracterización propia de este modelo continuo, donde coinciden las medidas de tendencia central, media, moda y mediana; la simetría respecto a estos parámetros y la facilidad de su aplicación hacen de la distribución normal, una herramienta de uso común, máxime que la mayoría de las variables económicas y sociales se ajustan a una función normal.

La distribución normal, también es útil como aproximación de los modelos binomial y poisson expuestos anteriormente, y yendo un poco más adelante, sustentados en el teorema del “límite central” podemos afirmar que, cuando el tamaño de la muestra es lo suficientemente grande, podemos asumir el supuesto de normalidad para una suma de

variables.

La forma acampanada de la variable normal, resalta la perfección de esta curva definida por los parámetros μ_x y σ_x^2

$$f(x) = \begin{cases} \dfrac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(x-u)^2}{2\sigma_x^2}} & , \quad para \quad -\infty < x < \infty \\ 0 & , \quad en \quad otro \quad caso \end{cases}$$

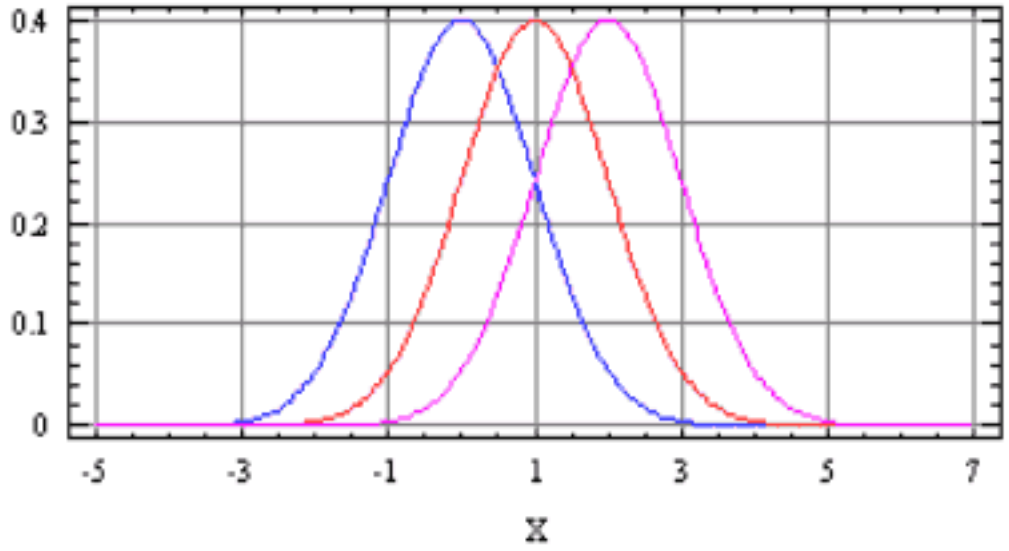
se representa como: $X \sim N(\mu_x, \sigma_x^2)$

La aparente complejidad de la distribución normal no debe preocupar al lector, donde:

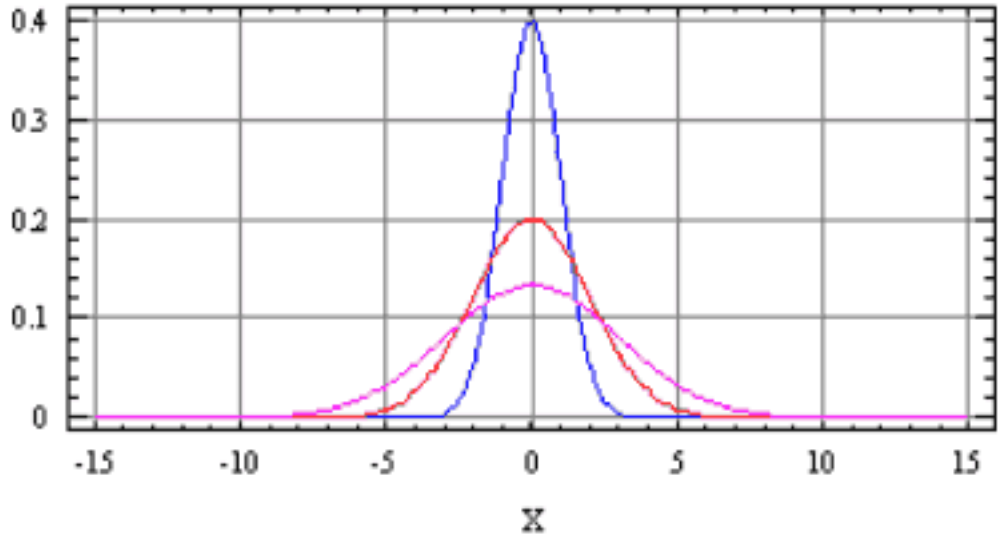
X : Variable aleatoria distribuida normalmente
μ_x Media aritmética de la variable
σ_x^2 Varianza de la variable
e 2.71828 constante (base de los logaritmos naturales)
π 3.1416 constante

Sin embargo, existen infinitas distribuciones normales, ya que por cada media aritmética ó varianza diferente se describe una función también diferente:

Normal Diferente Media Igual Varianza



Normal Diferente Varianza Igual Media


0.4
0.3
0.2
0.1
0
-15
-10
-5
0
5
10
15
X

11. Distribuciones Especiales

11.5 DISTRIBUCIÓN NORMAL ESTANDAR

Con el sinnúmero de diferentes distribuciones normales que se generarían con cada media o varianza diferente, se hace necesario efectuar un cambio de origen y de escala en la variable original, para estandarizarla y obtener una nueva variable cuya manipulación es más fácil:

$$Z = \frac{X - \mu_x}{\sigma_x},$$

con $Z \sim N(0,1)$, la nueva variable Z se distribuye normalmente con media aritmética $\mu_z = 0$ y varianza $\sigma_z^2 = 1$

$$f(z) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} & para \quad -\infty < z < \infty \\ 0 & en \quad otro \quad caso \end{cases}$$

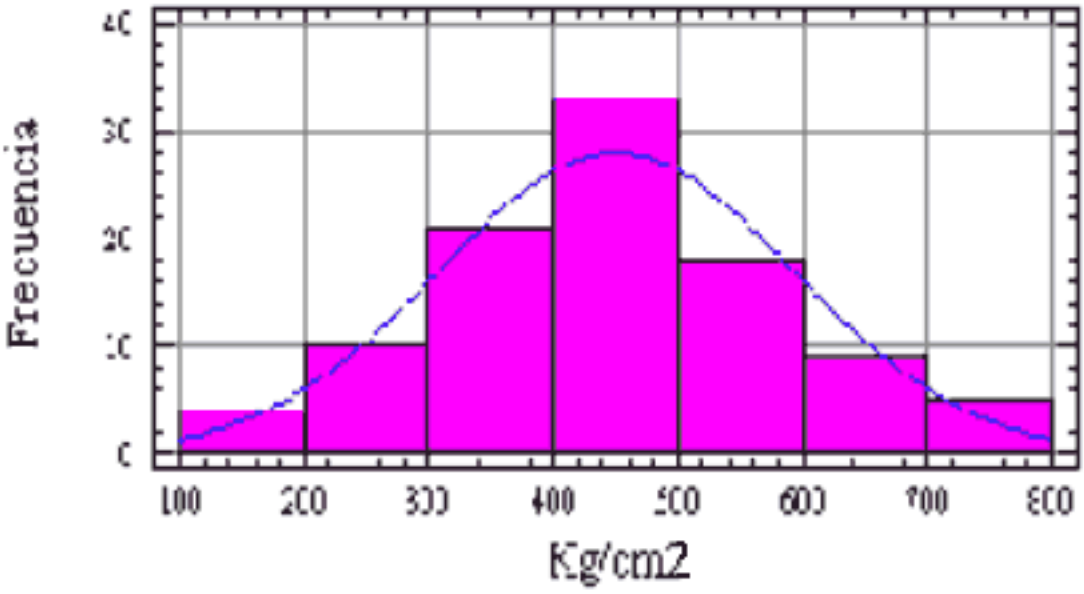
Dado que la distribución normal es una variable continua

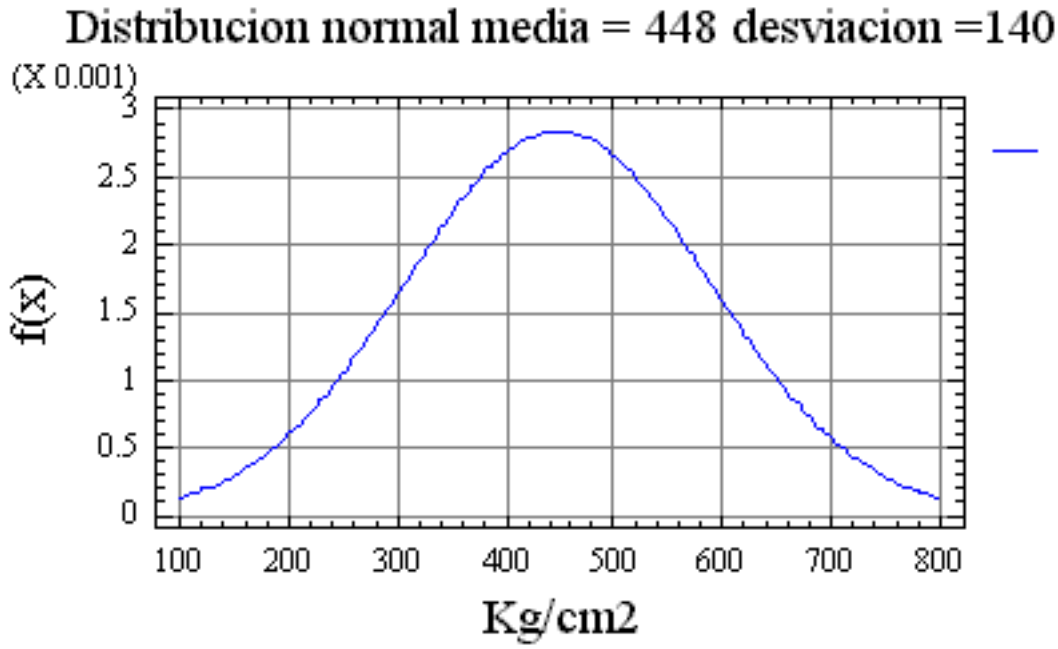
$$p(X = x) = 0 \quad , \quad p(X < x) = p(X \leq x)$$

Ejemplo:

Si asumimos que la resistencia de las baldosas se distribuye normalmente con $\mu_x = 448 kg/cm^2$ y $\sigma_x = 140 Kg/cm^2$

Resistencia de 100 Baldosas





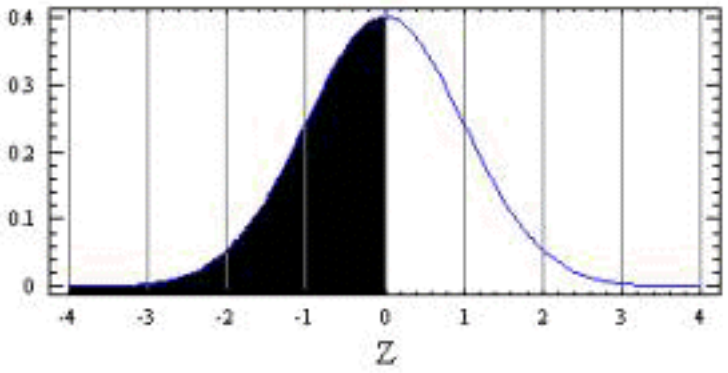
Si extraemos una baldosa al azar : Cual es la probabilidad de que:

1. Resista menos de 448 Kg/cm^2?
2. Resista más de 588 Kg/cm^2 ?
3. Resista entre 308 y 588 Kg/cm^2 ?
4. Resista entre 168 y 728 Kg/cm^2 ?
5. Resista más de 600 Kg/cm^2 ?
6. Resista menos de 200 ó más de 700 Kg/cm^2 ?

Con la ayuda de los valores tabulados:

1. $$p(X < 448) = p\left(\frac{X-\mu_x}{\sigma_x} \le \frac{448-448}{140}\right)$$

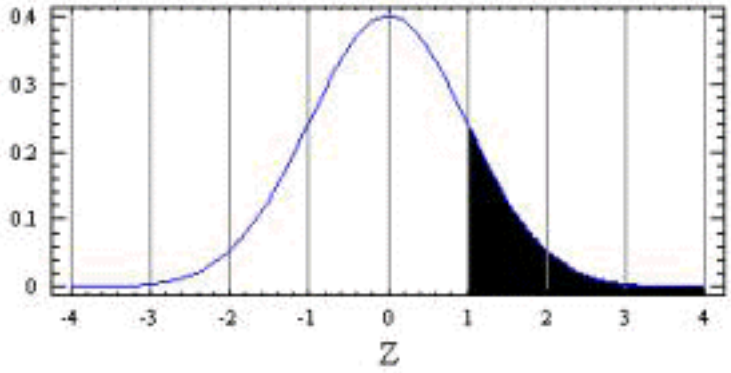
$$= p(Z \le 0) = 0.5$$



2. $$p(X > 588) = 1 - p(X \le 588)$$

$$= 1 - p\left(\frac{X-\mu_x}{\sigma_x} \le \frac{588-448}{140}\right)$$

$$= 1 - p(Z \le 1) = 1 - 0.8413 = 0.1587$$

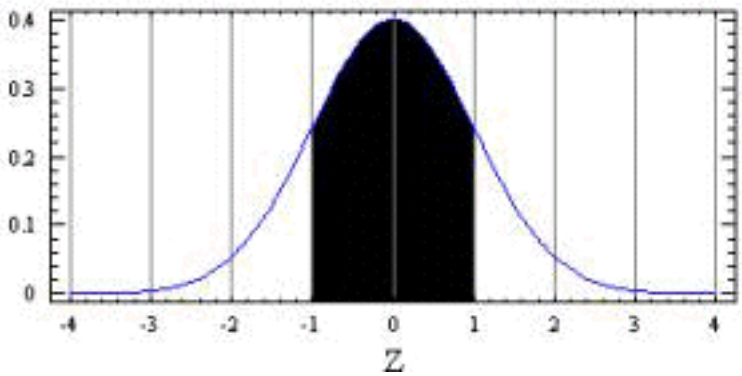


3. $$p(308 < X \le 588)$$

$$= p\left(\frac{308-448}{140} < \frac{X-\mu_x}{\sigma_x} \le \frac{588-448}{140}\right) = p(-1 < Z \le 1)$$

$$= p(Z \le 1) - p(Z \le -1) = p(Z \le 1) - \left[1 - p(Z \le 1)\right]$$

$$= 2p(Z \le 1) - 1 = 2(0.8413) - 1 = 0.6826$$



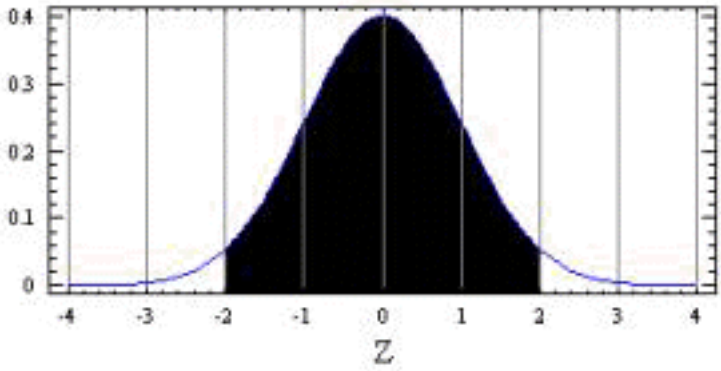
4. $p(168 < X \le 728)$

$$= p\left(\frac{168-448}{140} < \frac{X-\mu_x}{\sigma_x} \le \frac{728-448}{140}\right)$$

$$= p(-2 < Z \le 2) = p(Z \le 2) - p(Z \le -2)$$

$$= p(Z \le 2) - [1 - p(Z \le 2)]$$

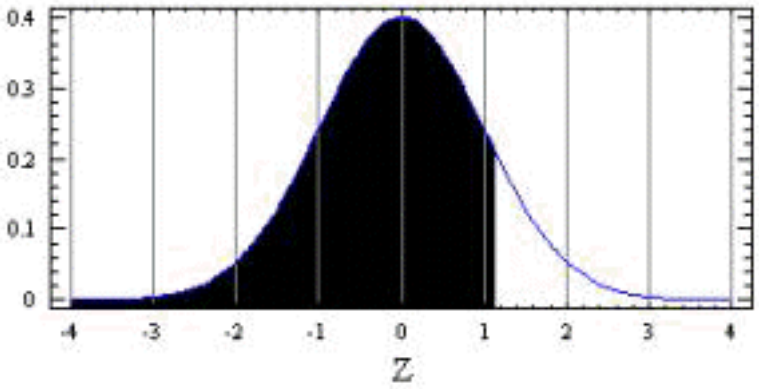
$$= 2p(Z \le 2) - 1 = 2(0.9772) - 1 = 0.9544$$



5. $p(X > 600) = 1 - p(X \le 600)$

$$= 1 - p\left(\frac{X-\mu_x}{\sigma_x} \le \frac{600-448}{140}\right)$$

$$= 1 - p(Z \le 1.08) = 1 - 0.86 = 0.14$$



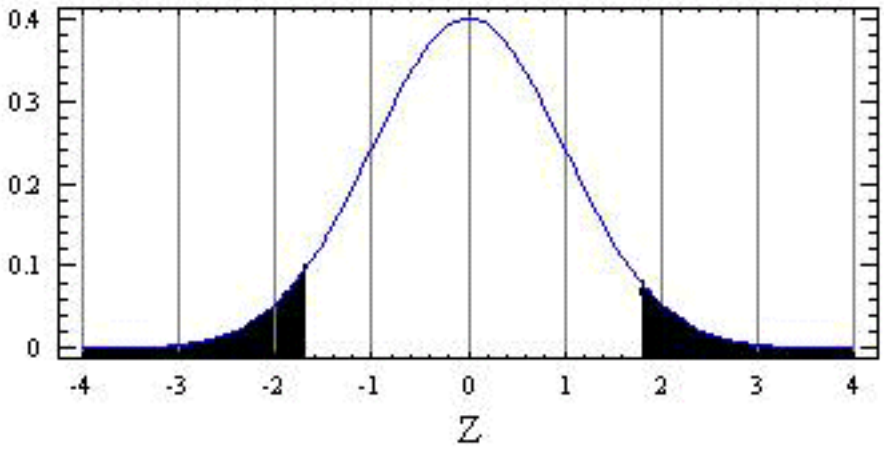
6. $p[(X < 200) \cup (X > 700)] = p(X < 200) + p(X > 700)$

$$= p\left(\frac{X-\mu_x}{\sigma_x} < \frac{200-448}{140}\right) + 1 - p\left(\frac{X-\mu_x}{\sigma_x} \le \frac{700-448}{140}\right)$$

$$= p(Z < -1.77) + 1 - p(Z \le 1.8)$$

$$= 1 - p(Z \le 177) + 1 - p(Z \le 1.8)$$

$$= 1 - 0.9616 + 1 - 0.9641 = 0.0743$$



$$Z \sim N(0,1) \quad \rightarrow \quad p(Z \leq z) = F(z) \qquad F(-z) = 1 - F(z)$$

Z	F(z)	z	F(z)	z	F(z)	z	F(z)
0	0.500	**0.82**	0.7939	**1.64**	0.9495	**2.46**	0.9931
0.02	0.508	**0.84**	0.7995	**1.66**	0.9515	**2.48**	0.9934
0.04	0.516	**0.86**	0.8051	**1.68**	0.9535	**2.5**	0.9938
0.06	0.524	**0.88**	0.8106	**1.7**	0.9554	**2.52**	0.9941
0.08	0.532	**0.9**	0.8159	**1.72**	0.9573	**2.54**	0.9945
0.1	0.540	**0.92**	0.8212	**1.74**	0.9591	**2.56**	0.9948
0.12	0.548	**0.94**	0.8264	**1.76**	0.9608	**2.58**	0.9951
0.14	0.556	**0.96**	0.8315	**1.78**	0.9625	**2.6**	0.9953
0.16	0.564	**0.98**	0.8365	**1.8**	0.9641	**2.62**	0.9956
0.18	0.571	**1**	0.8413	**1.82**	0.9656	**2.64**	0.9959
0.2	0.579	**1.02**	0.8461	**1.84**	0.9671	**2.66**	0.9961
0.22	0.587	**1.04**	0.8508	**1.86**	0.9686	**2.68**	0.9963
0.24	0.595	**1.06**	0.8554	**1.88**	0.9699	**2.7**	0.9965
0.26	0.603	**1.08**	0.8599	**1.9**	0.9713	**2.72**	0.9967
0.28	0.610	**1.1**	0.8643	**1.92**	0.9726	**2.74**	0.9969
0.3	0.618	**1.12**	0.8686	**1.94**	0.9738	**2.76**	0.9971
0.32	0.626	**1.14**	0.8729	**1.96**	0.9750	**2.78**	0.9973
0.34	0.633	**1.16**	0.8770	**1.98**	0.9761	**2.8**	0.9974
0.36	0.641	**1.18**	0.8810	**2**	0.9772	**2.82**	0.9976
0.38	0.648	**1.2**	0.8849	**2.02**	0.9783	**2.84**	0.9977
0.4	0.655	**1.22**	0.8888	**2.04**	0.9793	**2.86**	0.9979

0.42	*0.663*	**1.24**	*0.8925*	**2.06**	*0.9803*	**2.88**	*0.9980*
0.44	*0.670*	**1.26**	*0.8962*	**2.08**	*0.9812*	**2.9**	*0.9981*
0.46	*0.677*	**1.28**	*0.8997*	**2.1**	*0.9821*	**2.92**	*0.9982*
0.48	*0.684*	**1.3**	*0.9032*	**2.12**	*0.9830*	**2.94**	*0.9984*
0.5	*0.691*	**1.32**	*0.9066*	**2.14**	*0.9838*	**2.96**	*0.9985*
0.52	*0.698*	**1.34**	*0.9099*	**2.16**	*0.9846*	**2.98**	*0.9986*
0.54	*0.705*	**1.36**	*0.9131*	**2.18**	*0.9854*	**3**	*0.9987*
0.56	*0.712*	**1.38**	*0.9162*	**2.2**	*0.9861*	**3.02**	*0.9987*
0.58	*0.719*	**1.4**	*0.9192*	**2.22**	*0.9868*	**3.04**	*0.9988*
0.6	*0.726*	**1.42**	*0.9222*	**2.24**	*0.9875*	**VALORES**	
0.62	*0.732*	**1.44**	*0.9251*	**2.26**	*0.9881*	**ESPECIALES**	
0.64	*0.739*	**1.46**	*0.9279*	**2.28**	*0.9887*	**z**	***F(z)***
0.66	*0.745*	**1.48**	*0.9306*	**2.3**	*0.9893*	**0.842**	*0.8000*
0.68	*0.752*	**1.5**	*0.9332*	**2.32**	*0.9898*	**1.036**	*0.8500*
0.7	*0.758*	**1.52**	*0.9357*	**2.34**	*0.9904*	**1.282**	*0.9000*
0.72	*0.764*	**1.54**	*0.9382*	**2.36**	*0.9909*	**1.645**	*0.9500*
0.74	*0.770*	**1.56**	*0.9406*	**2.38**	*0.9913*	**1.960**	*0.9750*
0.76	*0.776*	**1.58**	*0.9429*	**2.4**	*0.9918*	**2.326**	*0.9900*
0.78	*0.782*	**1.6**	*0.9452*	**2.42**	*0.9922*	**2.576**	*0.9950*
0.8	*0.788*	**1.62**	*0.9474*	**2.44**	*0.9927*	**3.090**	*0.9990*

CUESTIONARIO Y EJERCICIOS PROPUESTOS

1. La probabilidad de que un visitante efectúe una compra en un almacén, durante un día dado es 0.8. Si al negocio entran 20 clientes, ¿cuál es la probabilidad de que el almacén realice:

1.1 Exactamente 16 ventas?
1.2 Menos de 17 ventas?
1.3 Más de 14 ventas?
1.4 Exactamente 5 ventas?
1.5 ¿Cuál es el número esperado de ventas?

2. Si un almacén tiene en promedio 5 ventas por hora. ¿Cual es la probabilidad de que en una hora determinada:

2.1 Haya exactamente 4 ventas?
2.2 Haya más de 3 ventas?
2.3 No se efectúen ventas?

3. Una de cada 10 personas mayores de 40 años de una comunidad, sufren de hipertensión. Se toma una muestra de 50 personas mayores de 40 años. Utilizando primero la distribución binomial y luego la aproximación a la distribución de Poisson, responder y comparar los resultados:

3.1 ¿Cuál es la probabilidad que haya más de 4 hipertensos?
3.2 ¿Cuál es la probabilidad que haya exactamente 5 hipertensos?

4. Un lote de arandelas tiene un diámetro normal con media 10 milímetros y desviación típica 0.5 milímetros. Se toma una arandela al azar. ¿Cuál es la probabilidad de que tenga un diámetro:

4.1 Superior a 10.5 milímetros?
4.2 Entre 9 y 11 milímetros?
4.3 Menos de 9 milímetros?

11. Distribuciones Especiales

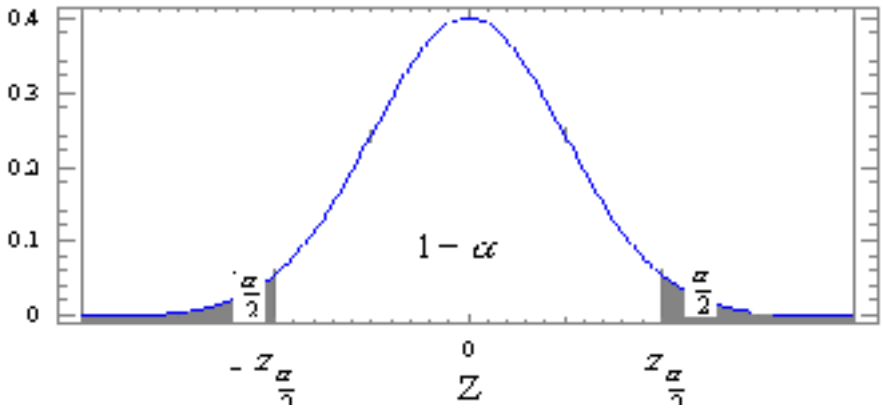
11.6 EL TAMAÑO DE LA MUESTRA

El teorema del limite central, sustenta la aproximación a la normalidad para muchas distribuciones discretas. Cuando el tamaño de la muestra es grande, y dicha muestra es tomada de cualquier distribución con media •, finita y varianza σ^2 finita, entonces la media aritmética muestral tiene una distribución normal con media μ y varianza $\frac{\sigma^2}{n}$

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}) \quad \rightarrow \quad Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Podemos entonces establecer intervalos de confianza para $\mu\, p(-k < Z \leq k) = probabilidad = nivel\ de\ confianza = 1 - \alpha$,

α : es denominado el nivel de significancia , si la significancia es por ejemplo, α=0.05 entonces la confiabilidad es del 0.95. $k = z_{\alpha/2}$



$$p(-k < Z < k) = p(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2})$$

$$= p(\mu - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} < \overline{X} < \mu + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}) = 1 - \alpha$$

Si $\alpha = 0.05$ entonces el 95% de las muestras se encontrarán en el intervalo

$$\left[\mu - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \mu + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right],$$

ahora bien , como los parámetros poblacional • y σ son desconocidos, para muestras grandes (*n>30)* la varianza muestral S^2 es un buen estimador de la varianza poblacional σ^2 , podemos afirmar con una confiabilidad predeterminada que la media aritmética poblacional • se halla en el intervalo

$$\left[\overline{X} - \frac{z_{\alpha/2}S}{\sqrt{n}}, \overline{X} + \frac{z_{\alpha/2}S}{\sqrt{n}}\right]$$

estamos admitiendo que la diferencia máxima entre • y $\overline{X}$ es de:

$$\frac{z_{\alpha/2}S}{\sqrt{n}}$$

esto es:

$$\overline{X} - \mu = e = \frac{z_{\alpha/2}S}{\sqrt{n}}$$

entonces el tamaño de muestra mínimo es

$$n = \frac{z_{\alpha/2}^2 S^2}{e^2}$$

donde :

$z_{\alpha/2}$: Valor crítico obtenido de la tabla normal, para una confiabilidad de $1-\alpha$

$S^{2:}$ Varianza muestral

e: Error máximo admitido

Sin embargo, *n* está en función de la varianza, la cual en la práctica es desconocida, ante lo cual debemos hacer un muestreo piloto para estimar la varianza y proceder a reajustar el tamaño de la muestra mínimo.

Ejemplo:

Se desea realizar una investigación para analizar, cual es la resistencia promedio de una producción de baldosas. Si admitimos un error máximo 25 Kg/Cm^2, cual debe ser el tamaño de muestra mínimo si exigimos una confiabilidad del 95%, y en una muestra piloto obtuvimos una desviación típica de 140 Kg/Cm^2?

$$n = \frac{z_{\alpha/2}^2 S^2}{e^2} = \frac{(1.96)^2 (140)^2}{(25)^2} = 121$$

Con una confiabilidad del 90% se quiere estimar la proporción de ciudadanos que votará en las próximas elecciones. Cual debe ser el tamaño de la muestra, si admitimos un error del 3% y se sabe que en las pasadas elecciones hubo una abstención del 70%?

Dado que X es una distribución binomial con $\bullet_x = np$ y $S_x^2 = np(1-p)$

Entonces

$$\overline{P} \sim N(p, \frac{p(1-p)}{n})$$

por consiguiente

$$\overline{P} - p = e = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

$$n = \frac{z_{\alpha/2}^2 p(1-p)}{e^2} = \frac{(1.645)^2 0.3(0.7)}{(0.03)^2} = 632$$

Tamaño de muestra confiabilidad = 80 %										
	p									
error	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.005	2041	3867	5478	6875	8057	9023	9775	10312	10635	10742
0.01	510	967	1370	1719	2014	2256	2444	2578	2659	2686
0.015	227	430	609	764	895	1003	1086	1146	1182	1194
0.02	128	242	342	430	504	564	611	645	665	671
0.025	82	155	219	275	322	361	391	412	425	430
0.03	57	107	152	191	224	251	272	286	295	298
0.035	42	79	112	140	164	184	199	210	217	219
0.04	32	60	86	107	126	141	153	161	166	168

Tamaño de muestra confiabilidad = 90 %										
	p									
error	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.005	5141	9740	13798	17316	20292	22727	24621	25973	26785	27056
0.01	1285	2435	3450	4329	5073	5682	6155	6493	6696	6764
0.015	571	1082	1533	1924	2255	2525	2736	2886	2976	3006
0.02	321	609	862	1082	1268	1420	1539	1623	1674	1691
0.025	206	390	552	693	812	909	985	1039	1071	1082
0.03	143	271	383	481	564	631	684	721	744	752
0.035	105	199	282	353	414	464	502	530	547	552
0.04	80	152	216	271	317	355	385	406	419	423
0.045	63	120	170	214	251	281	304	321	331	334
0.05	51	97	138	173	203	227	246	260	268	271
0.055	42	80	114	143	168	188	203	215	221	224
0.06	36	68	96	120	141	158	171	180	186	188
0.065	30	58	82	102	120	134	146	154	158	160

Tamaño de muestra confiabilidad = 95 %										
	P									
error	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.005	7299	13829	19592	24585	28811	32268	34957	36878	38031	38415
0.01	1825	3457	4898	6146	7203	8067	8739	9220	9508	9604
0.015	811	1537	2177	2732	3201	3585	3884	4098	4226	4268
0.02	456	864	1224	1537	1801	2017	2185	2305	2377	2401
0.025	292	553	784	983	1152	1291	1398	1475	1521	1537
0.03	203	384	544	683	800	896	971	1024	1056	1067
0.035	149	282	400	502	588	659	713	753	776	784
0.04	114	216	306	384	450	504	546	576	594	600
0.045	90	171	242	304	356	398	432	455	470	474
0.05	73	138	196	246	288	323	350	369	380	384
0.055	60	114	162	203	238	267	289	305	314	317
0.06	51	96	136	171	200	224	243	256	264	267
0.065	43	82	116	145	170	191	207	218	225	227
0.07	37	71	100	125	147	165	178	188	194	196
0.075	32	61	87	109	128	143	155	164	169	171
0.08	29	54	77	96	113	126	137	144	149	150

Tamaño de muestra confiabilidad = 98 %										
	p									
error	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.005	10283	19483	27601	34636	40589	45460	49248	51954	53578	54119
0.01	2571	4871	6900	8659	10147	11365	12312	12989	13394	13530
0.015	1143	2165	3067	3848	4510	5051	5472	5773	5953	6013
0.02	643	1218	1725	2165	2537	2841	3078	3247	3349	3382
0.025	411	779	1104	1385	1624	1818	1970	2078	2143	2165
0.03	286	541	767	962	1127	1263	1368	1443	1488	1503
0.035	210	398	563	707	828	928	1005	1060	1093	1104
0.04	161	304	431	541	634	710	770	812	837	846
0.045	127	241	341	428	501	561	608	641	661	668
0.05	103	195	276	346	406	455	492	520	536	541
0.055	85	161	228	286	335	376	407	429	443	447
0.06	71	135	192	241	282	316	342	361	372	376
0.065	61	115	163	205	240	269	291	307	317	320
0.07	52	99	141	177	207	232	251	265	273	276
0.075	46	87	123	154	180	202	219	231	238	241
0.08	40	76	108	135	159	178	192	203	209	211
0.085	36	67	96	120	140	157	170	180	185	187
0.09	32	60	85	107	125	140	152	160	165	167

Tamaño de muestra confiabilidad = 99%										
	P									
error	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.005	12606	23886	33838	42463	49762	55733	60378	63695	65686	66349
0.01	3152	5971	8459	10616	12440	13933	15094	15924	16421	16587
0.015	1401	2654	3760	4718	5529	6193	6709	7077	7298	7372
0.02	788	1493	2115	2654	3110	3483	3774	3981	4105	4147
0.025	504	955	1354	1699	1990	2229	2415	2548	2627	2654
0.03	350	663	940	1180	1382	1548	1677	1769	1825	1843
0.035	257	487	691	867	1016	1137	1232	1300	1341	1354
0.04	197	373	529	663	778	871	943	995	1026	1037
0.045	156	295	418	524	614	688	745	786	811	819
0.05	126	239	338	425	498	557	604	637	657	663
0.055	104	197	280	351	411	461	499	526	543	548
0.06	88	166	235	295	346	387	419	442	456	461
0.065	75	141	200	251	294	330	357	377	389	393
0.07	64	122	173	217	254	284	308	325	335	339
0.075	56	106	150	189	221	248	268	283	292	295
0.08	49	93	132	166	194	218	236	249	257	259
0.085	44	83	117	147	172	193	209	220	227	230
0.09	39	74	104	131	154	172	186	197	203	205
0.095	35	66	94	118	138	154	167	176	182	184
0.1	32	60	85	106	124	139	151	159	164	166

Apéndice No. 1

Una expresión general para cualquier media particular es la siguiente:

$$M(t) = \frac{(x_1^t f_1 + x_1^t f_2 + \ldots x_m^t f_m)^{1/t}}{f_1 + f_2 \ldots\ldots f_m}$$

Si consideramos que todos los datos tiene frecuencia igual a la unidad, es decir:

$$f_1 = f_2 = f_3 = \ldots\ldots = f_m = 1$$

La fórmula general queda:

$$M(t) = \frac{(x_1^t + x_2^t + \ldots\ldots x_n^t)^{1/t}}{n}$$

$$M(t) = \sqrt[t]{\frac{\sum x^t}{n}}$$

Donde:

M (t):	Media
Σ x :	Suma de todos los datos.
n:	Número de datos
t:	Un valor arbitrario -∞< t <∞

Como se puede observar; dándole valores a t, se obtienen medias particulares. Las más comunes son:

Para t = 1,

$$M(1) = \frac{(x_1^1 + x_2^1 + \ldots\ldots x_n^1)^{1/1}}{n} = \frac{\sum x}{n}$$

$$M(1) = \overline{X}, (la\ media\ aritmética)$$

Para t = -1,

$$M(-1) = \frac{(x_1^{-1} + x_2^{-1} + \ldots . x_n^{-1})^{-1}}{n}$$

$$M(-1) = \frac{\left[(\frac{1}{x_1} + \frac{1}{x_2} + \ldots . + \frac{1}{x_n}) \right]^{-1}}{n}$$

$$M(-1) = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots . + \frac{1}{x_n}}$$

La cual se denomina media armónica.

. . . . . Nótese que la media armónica es el recíproco de la media aritmética de los recíprocos de los datos.

Antiguamente la media armónica era llamada subcontraria.

Para t = 2,

$$M(2) = \frac{(x_1^2 + x_2^2 + \ldots . + x_n^2)^{1/2}}{n}$$

$$\overline{X_c} = \frac{(x_1^2 + x_2^2 + \ldots . + x_n^2)}{n}$$

$a\ \overline{X}_c = se\ le\ llama\ media\ cuadrática.$

Si t = 0,

$$M(0) = \frac{(x_1^0 + x_2^0 + \ldots . + x_n^0)^{1/0}}{n}$$

$$M(0) = \overline{X}_g; llamada\ media\ geométrica.$$

Para resolver la indeterminación presentada por: $\frac{1}{0} = \infty$

se procede de la siguiente forma:

$$\overline{X}_g = \frac{(x_1^t + x_2^t + x_3^t + \ldots + x_n^t)^{1/t}}{n} \; para \; t = 0$$

$$Log\ \overline{X}_g = \frac{1}{t} Log \frac{(x_1^t + x_2^t + \ldots + x_n^t}{n} \; para \; t = 0$$

$$Log\ \overline{X} = \lim_{t \to 0} \frac{1}{t} Log \frac{(x_1^t + x_2^t \ldots x_n^t)}{n}$$

Aplicando la regla de L' Hospital

$$\lim_{t \to 0} \log\ \overline{X} = \lim_{t \to 0} \frac{\dfrac{x_1^t \log(x_1) + x_2^t \log(x_2) + \ldots + x_n^t \log\ (x_n)}{n}}{\dfrac{x_1^t + x_2^t + \ldots + x_n^t}{n}}$$

$$Log\ \overline{X}_g = \frac{Log\ (x_1) + Log\ (x_2) + \ldots + Log\ (x_n)}{n}$$

$$Entonces,\ \overline{X}_g = Anti\log \frac{(Log\ x_1 + Log\ x_2 + \ldots + Log\ x_n}{n}$$

$$\overline{X}_g = (x_1 x_2 \ldots x_n)^{1/n}$$

$$\overline{X}_g = \sqrt[n]{x_1 . x_2 \ldots x_n .}$$

Apéndice No. 2

Para obtener la fórmula $m \cong +3.3\ Log(n)$ se deben hacer los siguientes supuestos:

1. El mínimo de datos que amerita clasificación en intervalos es 16.
2. El número de intervalos no debe ser inferior a (5).
3. Cada vez que se duplique la información se incrementa en uno (1) el número de intervalos.

Así las cosas, se obtiene la siguiente correspondencia:

Número de datos n	Número de intervalos m
$16 = 2^4$	5
$32 = 2^5$	6
$64 = 2^6$	7
$128 = 2^7$	8
$256 = 2^8$	9
.	.
.	.
.	.
$n = 2^{m-1}$	m

Se llega a la siguiente igualdad:

$$n = 2^{m} - 1$$

Tomando logaritmo a ambos lados de la ecuación

$$Log\ (n) = Log\ (2^{m-1}), \quad Log\ (n) = (m-1)\ Log(2)$$

$$1+\frac{Log\ (n)}{Log\ 2} = m \quad \rightarrow m = 1+\frac{Log\ (n)}{0.30103}$$

$$m = 1+(3.322)\ Log.\ (n)$$

Apéndice No. 3

Ejercicio general de aplicación.

Número de Personas por Familia y Consumo Diario de Arroz, Según Encuesta a 50 Familias de Quibdó

N° personas X	Consumo Y: gramos	N° personas X	Consumo Y: gramos
6	420	8	1080
8	890	7	660
5	520	5	580
6	770	5	730
10	1360	6	940
9	1280	6	470
6	590	7	710
2	80	4	380
5	670	6	500
7	670	8	1150
6	680	7	930
3	150	6	650
4	300	8	1120
4	320	3	240
8	1150	4	580
7	740	5	510
9	1300	4	470
6	720	6	570
6	870	8	1040
5	280	7	990
7	520	7	830
7	920	5	460
3	190	6	440
5	420	7	880
2	120	5	340

Como puede observarse, la variable X (número de personas por familia) es de tipo discreto, cuyos ítems no necesitan agruparse en intervalos; sin embargo, la variable Y (gramos de arroz diarios consumidos por familia) es una variable de tipo continuo y sus valores tienen poca frecuencia; en consecuencia, se hace necesario la agrupación en clases o intervalos:

1. $Y\max = 1360 \qquad Y\min = 80$

2. $R = Y\max - Y\min = 1360 - 80 = 1280$

3. $m = 1 + 3.3.\ Log\ (n) \quad m = 3.3\ Log\ (1.69)$

$m \cong 6.57 \quad m = 7$

$$A > \frac{R}{m} \quad , \quad A > \frac{1280}{7} \quad , \quad A > 182.8$$

$A = 200$

4. $Ra = mA \quad ; \quad Ra = (7)\ 200 = 1400$

5. $a = Ra - R \qquad a = 1400 - 1280 = 120$

6. $a \quad \left\{ \begin{matrix} Y\min - \cong a/2 = LIPI \\ Y\max + \cong a/2 = LSUI \end{matrix} \right\}$

$120 \left\{ \begin{matrix} 80 - 80 = 0 \\ 1360 + 40 = 1400 \end{matrix} \right\}$

7. Construcción de los intervalos:

Granos		
0	-	200
200	-	400
400	-	600
600	-	800
800	-	1000
1000	-	1200
1200	-	1400

Así las cosas, se procede a construir una tabla de doble entrada, con las variables X, Y para efectuar el conteo respectivo.

Numero de Personas por Familia
y Consumo Diario de Arroz, en un Grupo de 50 Familias

NUMERO DE PERSONAS POR FAMILIA															
CONSUMO GRAMOS DE ARROZ			X	2	3	4	5	6	7	8	9	10	fy	fay	fry	fray
	Intervalos	Y														
	1200-1400	1300									2	1	3	50	0.06	1.00
	1000- 1200	1100								5			5	47	0.10	0.94
	800-1000	900						2	5	1			8	42	0.16	0.84
	600-800	700					2	4	4				10	34	0.20	0.68
	400-600	500				2	5	6	1				14	24	0.28	0.48
	2000-400	300			1	3	2						6	10	0.12	0.20
	0-200	100		2	2								4	4	0.08	0.08
		fx		2	3	5	9	12	10	6	2	1	50		1.00	
		fax		2	5	10	19	31	41	47	49	50				
		frx		0.04	0.06	0.10	0.18	0.24	0.20	0.12	0.04	0.02	1.00			
		frax		0.04	0.10	0.20	0.38	0.62	0.82	0.94	0.98	1.0				

De la anterior tabla se pueden obtener, entre otras, las siguientes conclusiones:

- El 62 % de las familias las conforman entre 5 y 7 personas.
- El 62 % de las familias tienen 6 personas o menos.
- El 64 % de las familias consume entre 400 y 800 gramos de arroz, diariamente.
- El 84 % de las familias consume menos de 1000 gramos diarios de arroz.

Calculemos ahora las medidas de tendencia central y dispersión más importantes, a través de la siguiente tabla:

Tabla de Trabajo para el Cálculo de $\overline{X}, \overline{Y}, Sx, Sy$

	NUMERO DE PERSONAS POR FAMILIA															
CONSUMO DE ARROZ	**Intervalos**	Y / X	2	3	4	5	6	7	8	9	10	fy	yfy	y - 656	(y-656)**2	y - (656)**2fy
	1200-1400	1300								2	1	3	3900	644	414736	1244208
	1000-1200	1100							5			5	5500	444	197136	985680
	800-1000	900					2	5	1			8	7200	244	59536	476288
	600-800	700				2	4	4				10	7000	44	1936	19360
	400-600	500			2	5	6	1				14	7000	-156	24336	340704
	200-400	300		1	3	2						6	1800	-356	126736	760416
	0-200	100	2	2								4	400	-556	309136	1236544
	fx		2	3	5	9	12	10	6	2	1	50	32800			5063200
	Xfx		4	9	20	45	72	70	48	18	10	296				
	X - 5.92		-3.92	-2.92	-1.92	-0.92	0.08	1.08	2.08	3.08	4.08					
	(X-5.92)**2		15.37	8.53	3.69	0.85	0.01	1.17	4.33	9.49	16.65					
	(x -5.92)**2fx		30.73	25.58	18.43	7.62	0.08	11.66	25.96	18.97	16.65	155.68				

$$\overline{Y} = 656 \ gramos / día$$

$$\overline{X} = 5.92 \ personas / familia$$

$$S_x = \sqrt{S_x^2} = \sqrt{\frac{155.68}{50}}$$

$$S_x = 1.76$$

$$S_y = \sqrt{S_y^2} = \sqrt{\frac{5063200}{50}}$$

$$S_y = 318.2$$

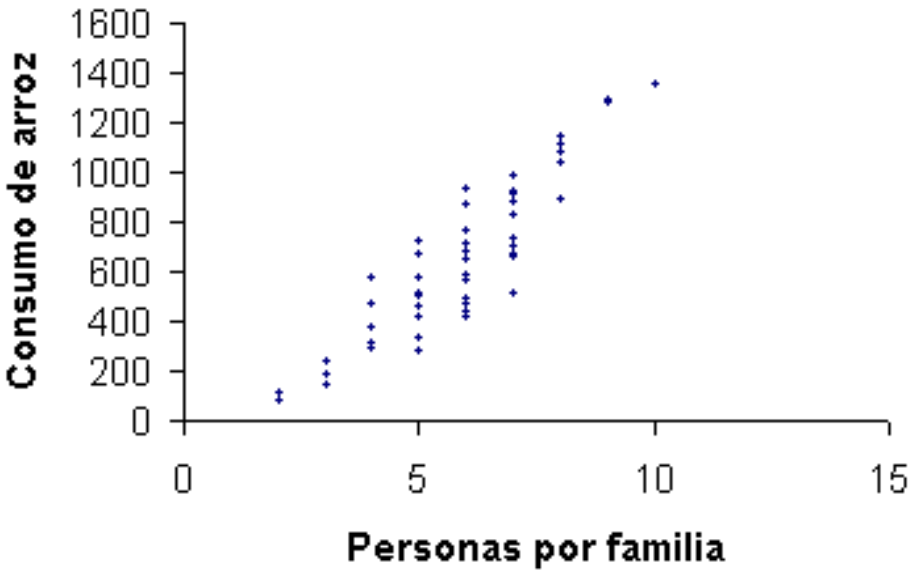
$$C_{vx} = \frac{S_x}{\overline{X}}(100) = 29.72\%$$

$$C_{vy} = \frac{S_y}{\overline{Y}}(100) = 48.5\%$$

Luego, es más dispersa la variable *y***:** Consumo de arroz.

Búsqueda de la correlación entre las dos variables.

1. Diagrama de Dispersión



Se vislumbra una correlación positiva y una tendencia rectilínea.

2. Cálculo del Coeficiente de Correlación.

Tabla de Trabajo para el Cálculo de Coeficiente de Correlacion

		PERSONAS POR FAMILIA											
Intervalos	Y / X	2	3	4	5	6	7	8	9	10	fy	yfy	Y**2fy
1200-1400	1300								2	1	3	3900	5070000
1000-1200	1100							5			5	5500	6050000
800-1000	900					2	5	1			8	7200	6480000
600-800	700				2	4	4				10	7000	4900000
400-600	500			2	5	6	1				14	7000	3500000
200-400	300		1	3	2						6	1800	540000
0-200	100	2	2								4	400	40000
Fx		2	3	5	9	12	10	6	2	1	50	32800	26580000
Xfx		4	9	20	45	72	70	48	18	10	296		
X**2fx		8	27	80	225	432	490	384	162	100	1908		

Debido a que los datos están agrupados en frecuencias, debemos hacer los cálculos teniendo en cuenta las veces que cada par de observaciones se repite.

$$xyf_{xy} = 2(100)2 + 3(100)2 + 3(300)1 + 4(300)3 + 4(500)2$$
$$+ 5(300)2 + 5(500)5 + 5(700)2 + 6(500)6 + 6(700)4$$
$$+ 6(900)2 + 7(500)1 + 7(700)4 + 7(900)5 + 8(900)1$$
$$+ 8(1100)5 + 9(1300)2 + 10(1300)1$$

$$r = \frac{n\sum xyf_{xy} - (\sum xf_x)(\sum xf_y)}{\sqrt{\left[n\sum x^2 f_x - (\sum xf_x)^2\right]\left[n\sum y^2 f_y - (\sum yf_y)^2\right]}}$$

$$r = \frac{50(219800) - (296)(32800)}{\sqrt{\left[50(1908) - (296)^2\right]\left[50(26580000 - (32800)^2\right]}}$$

$$r = \frac{10990000 - 9708800}{\sqrt{(95400 - 87616)(1329000000 - 1075840000)}}$$

$$r = \frac{1281200}{1403779.7} = 0.91$$

Entre el número de personas por familia y el consumo de arroz, existe una correlación positiva excelente.

Se pueden utilizar los mismos totales para efectos de calcular los parámetros a y b de la recta de regresión.

$$Y = a + bx$$

$$b = \frac{n\sum xyf_{xy} - (\sum xf_x)(\sum yf_y)}{n\sum x^2 f_x - (\sum xf_x)^2}$$

$$a = \frac{\sum yf_y - b\sum xf_x}{n}$$

$$b = \frac{50(219800) - 296(32800)}{50(1908) - (296)^2}$$

$$b = \frac{12812}{7784} = 1.65$$

$$b = 1.65$$

$$a = \frac{328 - 1.65(296)}{50}$$

$$a = -3.2$$

El modelo lineal queda por consiguiente:

$$\hat{Y} = -3.2 + 1.65x$$

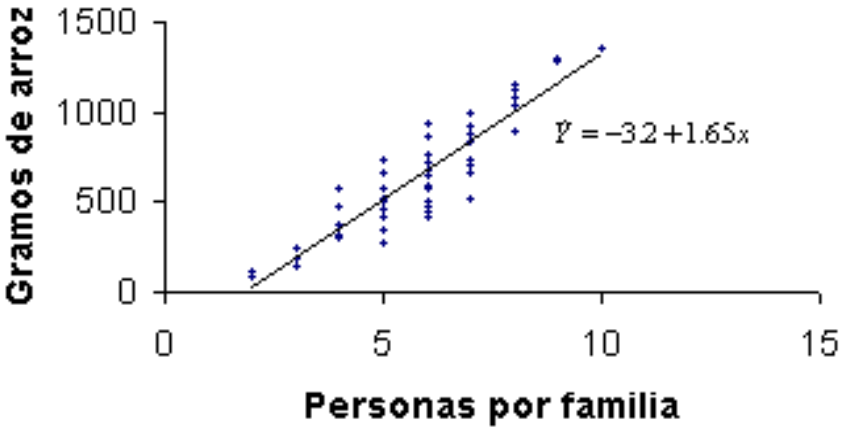
Modelo que predice el consumo de arroz, en función del número de personas por familia:

x: *Número de personas por familia*

$\hat{Y}$: *Estimación del consumo de arroz*

Para finalizar, veamos el comportamiento gráfico del modelo rectilíneo entre el número de personas por familia y el consumo de arroz.

Recta de Regresión


Gramos de arroz
1500
1000
500
0
0
5
10
15
Personas por familia
Y = −3.2 + 1.65x

Solución a Algunos Ejercicios Propuestos

3.1-6

PALABRAS POR MINUTO ESCRITAS POR UN GRUPO DE MECANOGRAFAS

Class	PALABRAS	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	75	12	0.0600	12	0.0600
2	76	17	0.0850	29	0.1450
3	77	25	0.1250	54	0.2700
4	78	26	0.1300	80	0.4000
5	79	29	0.1450	109	0.5450
6	80	33	0.1650	142	0.7100
7	81	22	0.1100	164	0.8200
8	82	19	0.0950	183	0.9150
9	83	11	0.0550	194	0.9700
10	84	6	0.0300	200	1.0000

3.1-7 Edad

EDAD DE 50 OPERARAIAS DE LA FABRICA DE CONFECCIONES LA HILACHA

Class	EDAD	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	22	1	0.0200	1	0.0200
2	23	1	0.0200	2	0.0400
3	24	3	0.0600	5	0.1000
4	25	4	0.0800	9	0.1800
5	26	4	0.0800	13	0.2600
6	27	5	0.1000	18	0.3600
7	28	6	0.1200	24	0.4800
8	29	4	0.0800	28	0.5600
9	30	4	0.0800	32	0.6400
10	31	3	0.0600	35	0.7000
11	32	2	0.0400	37	0.7400
12	33	2	0.0400	39	0.7800
13	35	3	0.0600	42	0.8400
14	36	2	0.0400	44	0.8800
15	38	3	0.0600	47	0.9400
16	39	2	0.0400	49	0.9800
17	40	1	0.0200	50	1.0000

3.1-7 Estado Civil

ESTADO CIVIL DE 50 OBRERAS EN LA FABRICA DE CONFECCIONES LA HILACHA

Class	ESTADO CIVIL	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	casada	10	0.2000	10	0.2000
2	soltera	30	0.6000	40	0.8000
3	u. Libre	5	0.1000	45	0.9000
4	viuda	5	0.1000	50	1.0000

3.1-7 Número de Hijos

NÚMERO DE HIJOS DE 50 OBRERAS DE LA FABRICA DE CONFECCIONES LA HILACHA

Class	No HIJOS	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	0	11	0.2200	11	0.2200
2	1	12	0.2400	23	0.4600
3	2	17	0.3400	40	0.8000
4	3	9	0.1800	49	0.9800
5	4	1	0.0200	50	1.0000

3.1-7 Experiencia Laboral

EXPERIENCIA LABORAL DE 50 OBRERAS DE LA FABRICA DE CONFECCIONES LA HILACHA

Class	Años de Experiencia	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	2	1	0.0200	1	0.0200
2	3	3	0.0600	4	0.0800
3	4	5	0.1000	9	0.1800
4	5	10	0.2000	19	0.3800
5	6	13	0.2600	32	0.6400
6	7	11	0.2200	43	0.8600
7	8	5	0.1000	48	0.9600
8	9	2	0.0400	50	1.0000

3.1-7 Escolaridad

ESCOLARIDAD DE 50 OBRERAS DE LA FABRICA DE CONFECCIONES LA HILACHA

Class	Años de Estudio	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	3	13	0.2600	13	0.2600
2	4	10	0.2000	23	0.4600
3	5	10	0.2000	33	0.6600
4	6	5	0.1000	38	0.7600
5	7	2	0.0400	40	0.8000
6	8	5	0.1000	45	0.9000
7	9	5	0.1000	50	1.0000

3.1-7 Gastos en Educación

GASTOS EN EDUCACION DE 50 OBRERAS EN LA FABRICA LA HILACHA (Miles $/mes)

Class	Miles $/mes	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	0	1	0.0200	1	0.0200
2	1	3	0.0600	4	0.0800
3	2	5	0.1000	9	0.1800
4	3	10	0.2000	19	0.3800
5	4	5	0.1000	24	0.4800
6	5	4	0.0800	28	0.5600
7	6	7	0.1400	35	0.7000
8	7	4	0.0800	39	0.7800
9	8	5	0.1000	44	0.8800
10	9	3	0.0600	47	0.9400
11	10	1	0.0200	48	0.9600
12	11	1	0.0200	49	0.9800
13	13	1	0.0200	50	1.0000

3.1-7 Ausencias

AUSENCIAS AL TRABAJO, DURANTE EL ULTIMO AÑO, DE 50 OBRERAS EN LA HILACHA

Class	Ausencias	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	0	10	0.2000	10	0.2000
2	1	12	0.2400	22	0.4400
3	2	15	0.3000	37	0.7400
4	3	7	0.1400	44	0.8800
5	4	4	0.0800	48	0.9600
6	5	2	0.0400	50	1.0000

3.2-4

CONSUMO DE AGUA, DURANTE OCTUBRE EN UN BARRIO RESIDENCIAL DE UNA CIUDAD (M3)

Class*	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	0.0	4.0	2.0	8	0.0435	8	0.0435
2	4.0	8.0	6.0	17	0.0924	25	0.1359
3	8.0	12.0	10.0	29	0.1576	54	0.2935
4	12.0	16.0	14.0	37	0.2011	91	0.4946
5	16.0	20.0	18.0	34	0.1848	125	0.6793
6	20.0	24.0	22.0	30	0.1630	155	0.8424
7	24.0	28.0	26.0	17	0.0924	172	0.9348
8	28.0	32.0	30.0	9	0.0489	181	0.9837
9	32.0	36.0	34.0	3	0.0163	184	1.0000

* No incluye el limite inferior

3.1-7 Calificación

CALIFICACION DEL REND IMIENTO LABORAL DE 50 OPERARIAS DE "LA HILACHA"

Class	Calificación	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	1	13	0.2600	13	0.2600
2	2	11	0.2200	24	0.4800
3	3	16	0.3200	40	0.8000
4	4	8	0.1600	48	0.9600
5	5	2	0.0400	50	1.0000

3.2-4

CONSUMO DE AGUA, DURANTE OCTUBRE EN UN BARRIO RESIDENCIAL DE UNA CIUDAD (M3)

Class*	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	0.0	4.0	2.0	8	0.0435	8	0.0435
2	4.0	8.0	6.0	17	0.0924	25	0.1359
3	8.0	12.0	10.0	29	0.1576	54	0.2935
4	12.0	16.0	14.0	37	0.2011	91	0.4946
5	16.0	20.0	18.0	34	0.1848	125	0.6793
6	20.0	24.0	22.0	30	0.1630	155	0.8424
7	24.0	28.0	26.0	17	0.0924	172	0.9348
8	28.0	32.0	30.0	9	0.0489	181	0.9837
9	32.0	36.0	34.0	3	0.0163	184	1.0000

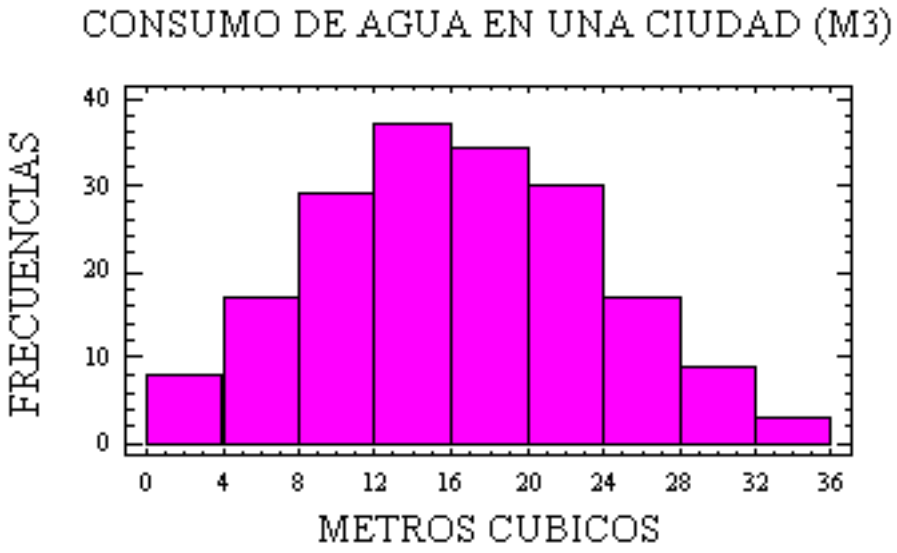
* No incluye el limite inferior

3.1-7 Calificación

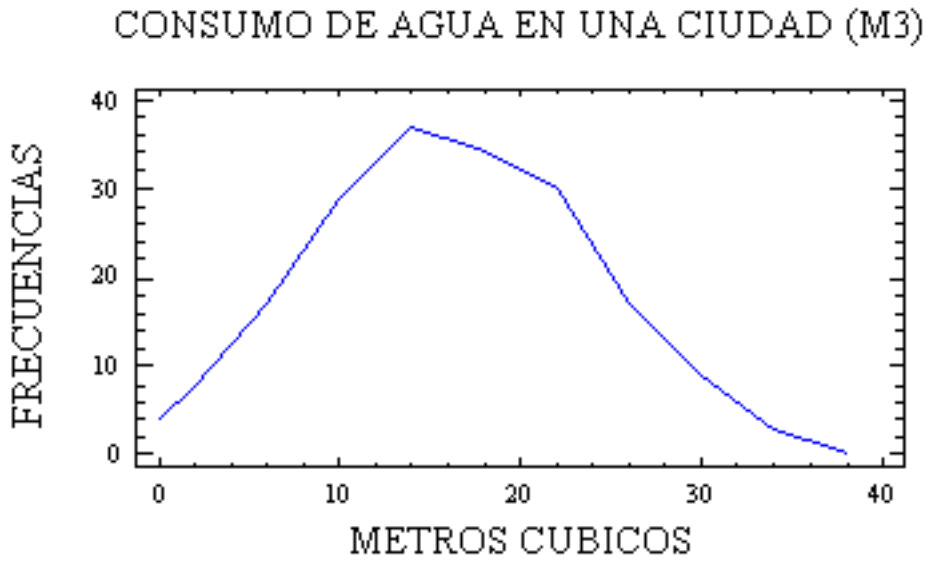
CALIFICACION DEL RENDIMIENTO LABORAL DE 50 OPERARIAS DE "LA HILACHA"

Class	Calificación	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	1	13	0.2600	13	0.2600
2	2	11	0.2200	24	0.4800
3	3	16	0.3200	40	0.8000
4	4	8	0.1600	48	0.9600
5	5	2	0.0400	50	1.0000

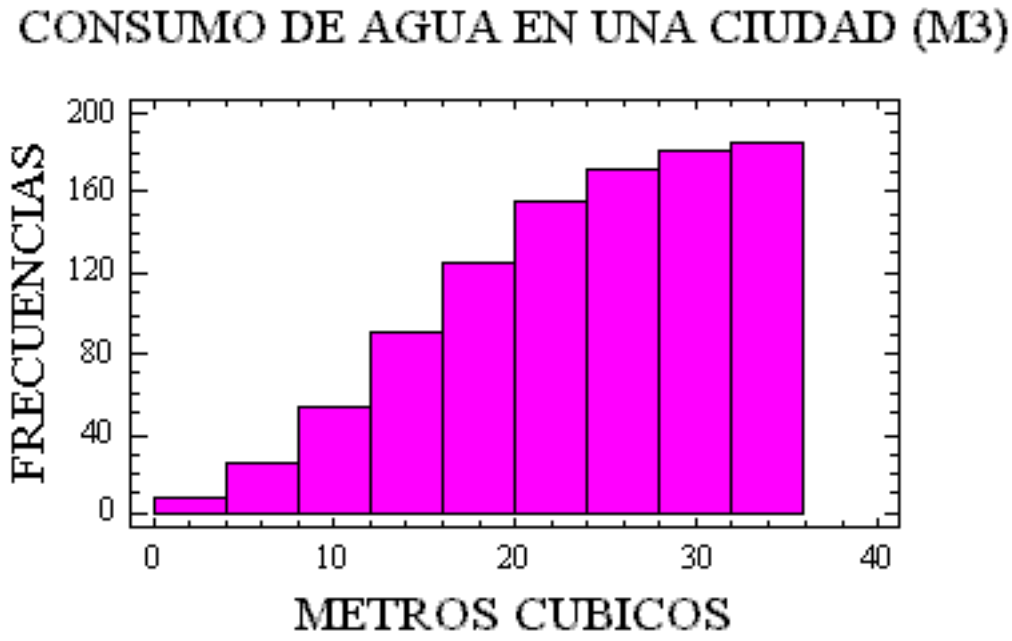
4-5 Histograma



4-5 Polígono

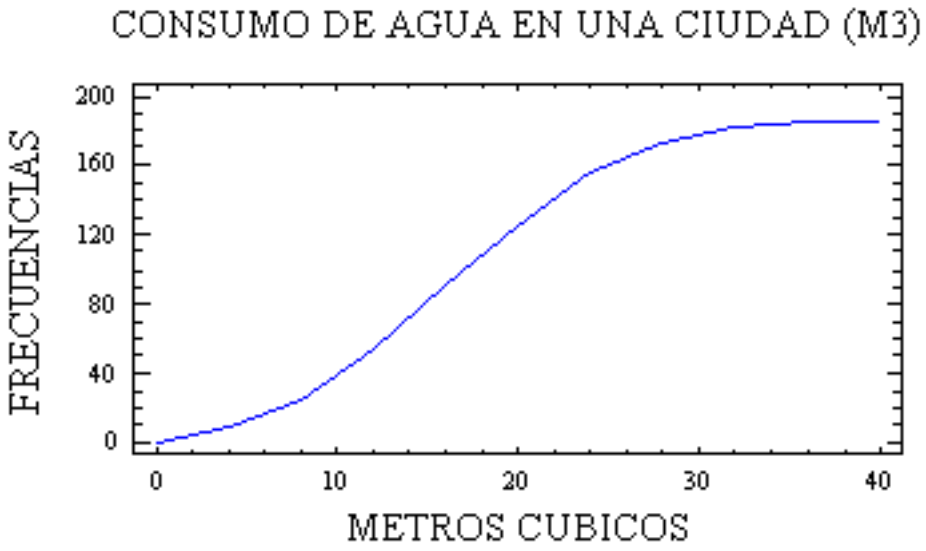


4-5 Histograma Acumulado

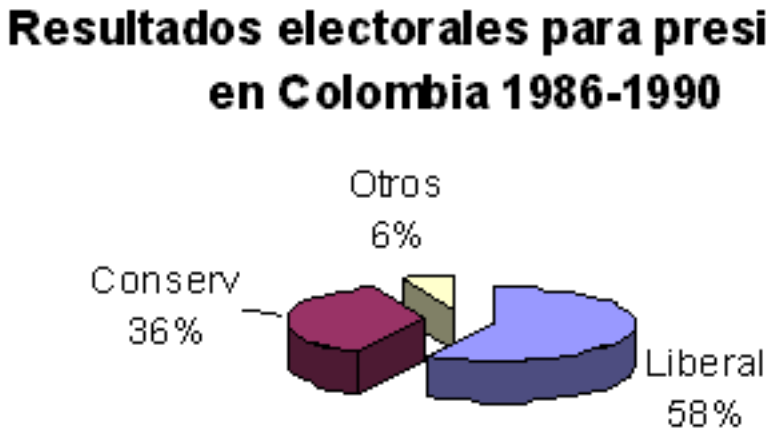


. . . .
.

4-5 Polígono de Frecuencias Acumuladas



4-8



5-6

	Palabras por minuto	Consumo agua m3	Edad años	Años experiencia	Años escolaridad
Count	50	50	50	50	50
Average	79.08	16.7	29.84	5.88	5.16
Median	79.0	17.0	29.0	6.0	5.0
Mode	78.0	17.0	28.0	6.0	3.0

	Gastos Educacion	calificacion	No. de hijos
Count	50	50	50
Average	5.08	2.5	1.54
Median	5.0	3.0	2.0
Mode	3.0	3.0	2.0

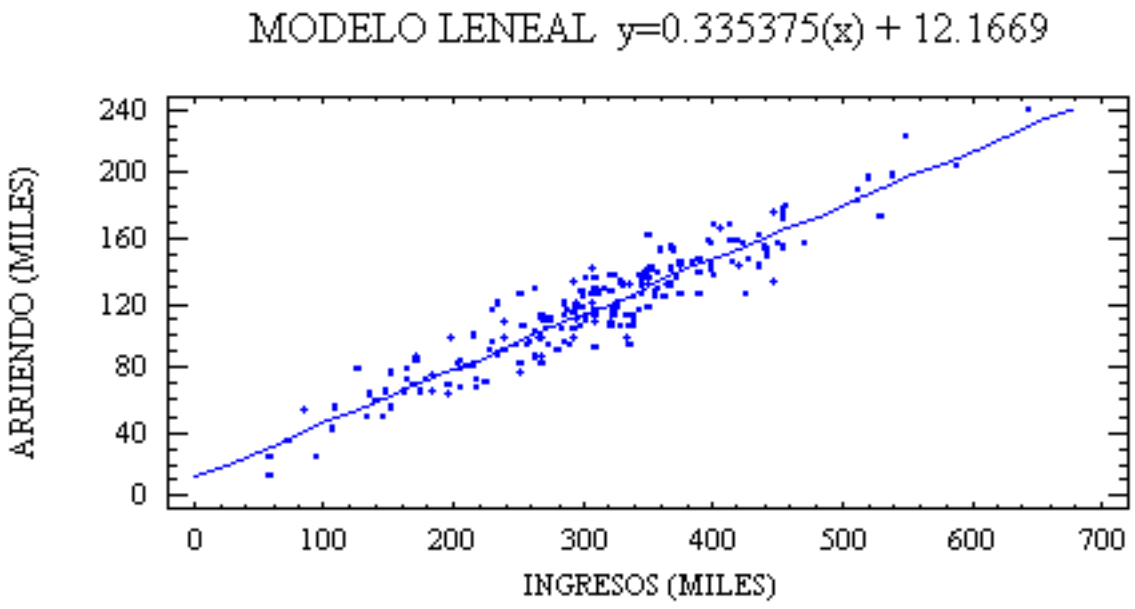
7-4

	Palabras por minuto	Salario/dia miles $	Consumo H2o m3	Calificacion	Años experiencia
Count	50	50	50	50	50
Variance	5.74857	3.27551	54.4592	1.35714	2.4751
Standard deviation	2.39762	1.80984	7.37965	1.16496	1.57325
Range	9.0	8.0	30.0	4.0	7.0
Coeff. of variation	3.03189%	3.34536%	44.1895%	46.5986%	26.7559%

	escolaridad	Gastos Educacion	Edad	No. de hijos	ausencias
Count	50	50	50	50	50
Variance	4.05551	8.23837	22.178	1.19224	1.84857
Standard deviation	2.01383	2.87026	4.70935	1.0919	1.35962
Range	6.0	13.0	18.0	4.0	5.0
Coeff. of variation	39.0277%	56.5011%	15.782%	70.9026%	76.3832%

8-1

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: arriendo
Independent variable: ingreso
-----------------------------------------------------------------------------
                                    Standard          T
Parameter          Estimate           Error       Statistic         P-Value
-----------------------------------------------------------------------------
Intercept           12.1669          2.7635         4.40273          0.0000
Slope              0.335375      0.00850132         39.4498          0.0000
-----------------------------------------------------------------------------

                              Analysis of Variance
-----------------------------------------------------------------------------
Source               Sum of Squares     Df  Mean Square     F-Ratio
-----------------------------------------------------------------------------
Model                      237781.0      1     237781.0     1556.29
Residual                    30251.8    198      152.787
-----------------------------------------------------------------------------
Total (Corr.)              268032.0    199

Correlation Coefficient = 0.941878
R-squared = 88.7134 percent


   arriendo = 12.1669 + 0.335375*ingreso

si ingreso = 270   entonces:   arriendo=12.1669+0.335375(270) =102.718

se estima debe pagar 102.718 pesos
```



9-7

					valor de q0			
	Cant	precios			a los precios de			
	q0	p0	p1	p2	p3	p0	p1	p2	p3
Art Año	1998	1998	1999	2000	2001	1998	1999	2000	2001
A	180	200	250	300	350	1000	1250	1500	1750
B	100	50	60	70	80	500	600	700	800
C	400	100	120	130	150	1500	1800	1950	2250
D	120	20	30	30	40	400	600	600	800
sumas						3400	4250	4750	5600
Indice						**1.00**	**1.25**	**1.40**	**1.65**

9-8

Año	MES	Valor Indice	Salario nominal	Salario Real
2000	1	110.64	260000	234996
2000	2	113.19	260000	229702
2000	3	115.12	260000	225851
2000	4	116.27	260000	223617
2000	5	116.88	260000	222450
2000	6	116.85	260000	222507
2000	7	116.81	260000	222584
2000	8	117.18	260000	221881
2000	9	117.68	260000	220938
2000	10	117.86	260000	220601
2000	11	118.24	260000	219892
2000	12	118.79	260000	218874
2001	1	120.04	286000	238254
2001	2	122.31	286000	233832
2001	3	124.12	286000	230422
2001	4	125.54	286000	227816
2001	5	126.07	286000	226858
2001	6	126.12	286000	226768

10.1-1 Si se usa el pulgar:

$$V_2^5 = \frac{5!}{3!} = 20$$

10.1-2 Variaciones con repetición

$$VR_5^6 = 6^5$$

10.1-5 Permutaciones con repetición

$$PR_{(4,2)}^6 = \frac{6!}{4!2!}$$

10.1-8 Combinaciones

$$C_2^4 = \frac{4!}{2!2!}$$

10.1-10 Combinaciones

$$C_2^x = \frac{x!}{(x-2)!2!} = 21$$

$$\frac{x(x-1)(x-2)!}{(x-2)!} = 42 \quad \rightarrow \quad x^2 - x - 42 = 0$$

$$(x-7)(x+6) = 0 \quad \rightarrow \quad x = 7$$

10.4-4

$$p(nV / M) = \frac{p(nV \cap M)}{p(M)} = \frac{40}{180}$$

10.6-2.1

$$p(X \leq 5) = \frac{27}{28}$$

10.6-2.2

$$p(X > 2) = 1 - p(X \leq 2) = 1 - \frac{18}{28} = \frac{10}{28}$$

10.6-2.3

$$p(2 < X \le 5) = p(X \le 5) - p(X \le 2) = \frac{27}{28} - \frac{18}{28} = \frac{9}{28}$$

10.6-2.4

$$p\left[(X > 5) \cup (X < 3)\right] = p(X > 5) + p(X < 3)$$
$$= 1 - p(x \le 5) + p(x \le 2) = 1 - \frac{27}{28} + \frac{18}{28} = \frac{19}{28}$$

10.5-4.1

$$p(X > 10.5) = p(\frac{X - \mu}{\sigma} > \frac{10.5 - 10}{0.5}) = p(Z > 1)$$
$$= 1 - p(Z \le 1) = 1 - 0.8413 = 0.1587$$

10.5-4.2

$$p(9 < X \le 11) = p(\frac{9 - 10}{0.5} < \frac{X - \mu}{\sigma} \le \frac{11 - 10}{0.5})$$
$$= p(-2 < Z \le 2) = p(Z \le 2) - p(Z \le -2)$$
$$= p(Z \le 2) - \left[1 - p(Z \le 2)\right] = 2p(Z \le 2) - 1$$
$$= 2(0.9772) - 1 = 0.9544$$

10.5-4.3

$$p(X < 9) = p(\frac{X - \mu}{\sigma} > \frac{9 - 10}{0.5}) = p(Z > -2)$$
$$= 1 - p(Z \le -2) = 1 - \left[1 - p(Z \le 2)\right] = p(Z \le 2)$$
$$= 0.9772$$

Referencias Bibliográficas

- Alatorre, *et al.*, *Introducción a los métodos estadísticos*, México, UPN.
- Azorín, Poch. Francisco. *Curso de muestreo y aplicaciones*, Aguilar, 1969.
- Barahoma, Abel y otro. *Metodología de trabajos científicos*, Ipler, 1979.
- Bencardino M., Ciro. *Estadística, Apuntes y 600 Problemas Resueltos*, 2ª Edición, Ecoe, 1982.
- Castillo, Juana, *Estadística inferencial básica*, México, CCH, UNAM.
- CHAO. Lincoln L. *Estadística para Ciencias Administrativas*, 2ª Edición, MCGRAW-HILL, 1980.
- Dixon, Wilfrid J y otro. *Introducción al Análisis Estadístico*, 2ª Edición, MCGRAW-HILL, 1965.
- Doms, Fernan P. *La Estadística Qué Sencilla*, 2ª Edición, Paraninfo, 1969.
- Downie, N. M. y otro. *Métodos Estadísticos Aplicados*. Harper Row Publishers Inc., 1970.
- Giardina, Basilio. *Manual de Estadística*, 3 Edición, 1972.
- Haber, Audrey. *Estadística General*, Fondo Educativo Interamericano, 1973.
- Hoel, Paul G. *Estadística elemental*, México, CECSA.
- Johnson, Robert, *Estadística elemental*, Buenos Aires, Grupo Editorial Iberoamericana.
- Kazmier, Leonard J. *Estadística Aplicada a la Administración y la Economía*, MCGRAW-HILL, 1978.
- Leitold Louis. *El Cálculo con Geometría Analítica*, 2ª Edición, Harla S.A., 1973.
- Levin Yack. *Fundamentos de Estadística en la Investigación Social*, 2ª Edición, Harla S., 1977.
- Llerena, León, Ricardo y otro. *Curso de Estadística General*, U. de A., 1981.

- Mejía V., William. *Bioestadística General,* Escuela Nal. De Salud Pública, U. de A., 1980.
- National Council of Teachers. Of. Mathematics USA. *Recopilación, Organización e interpretación de Datos,* Trilla, 1970.
- Portilla, Ch. Enrique. *Estadística, Primer Curso.* Interamericano, 1980.
- Richards, Larry E. Y otro. *Estadística en los Negocios. ¿porqué y cuándo?,* MCGRAW-HILL,1978.
- Seymour, Lipschutz, *Teoría y problemas de probabilidad*, México, McGraw-Hill.
- Shao, Stephen P. *Estadística para Economistas y Administradores de Empresas,* 15ª Edición, 1979.
- Spiegel, Murray R. *Estadística,* MCGRAW-HILL, 1970.
- Spiegel, Murray, *Teoría y problemas de estadística,* México, McGraw-Hill.
- Stevenson, William, *Estadística*, México, Harla.
- Yamane, Taro, *Estadística*, México, Harla.

NORBERTO GUARÍN SALAZAR

Estadístico Universidad de Medellín
M.Sc. Estadística Universidad Nacional de Colombia
Profesor Asociado Universidad del Chocó

. . . . .